THE UNIVERSITY
of EDINBURGH

Doctoral Thesis

Detecting Selection In The Evolution Of Cancer Genomes

Author:
Joanna Margaret Pethick

Supervisor:
Dr. Martin Taylor

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

The University Of Edinburgh

October 2015

Declaration of Authorship

I, Joanna Margaret Pethick, declare that this thesis titled, 'Detecting Selection In The Evolution Of Cancer Genomes' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Abstract

Cancer is a disease of the genome, requiring mutation or epimutation of specific genes to develop. The subsequent progression of cancer and response to therapies is also dictated to some degree by new mutation and clonal selection on that novel variation. However, it is thought that the majority of somatic mutations that occur in cancer are inconsequential passengers, and only a subset of functionally important driver mutations are of importance for cancer biology. This project set out to adapt and apply well-established methods from the field of molecular evolution to measure the selective forces driving the development of cancers. The ultimate objective being an improved understanding of which mutations help or hinder the progression of a cancer. Somatic cancer mutations were identified through the analysis of paired tumour and non-tumour whole-exome sequence data from the same individual. Primary data from 1005 patients was processed and complemented with additional publicly available pre-processed somatic variant calls from 4728 patients. Tumours were classified by tissue of origin and also their spectrum of substitution mutations. An advanced evolutionary analysis framework was established, allowing somatic single nucleotide variant data to be analysed as traditional organismal DNA sequence. Estimates of amino acid changing (non-synonymous) and synonymous mutation rates were derived and maximum likelihood tests of selection applied to identify genes and regions of genes subject to selective pressure during oncogenesis. While the meta-analysis of all patients provided unprecedented power for such a study, more refined analyses based on the stratification of patients gave insights into the pathways of importance for specific tissues of origin. Additionally, stratification of patients by the relative frequencies of different mutation types in a tumour also provided insights into how mutation profile influences the sites, genes and pathways that are perturbed in the development of cancer. Of particular interest here, was to test the hypothesis that both (1.) mutation spectrum and (2.) tissue of origin, set the evolutionary trajectory of a cancer. Building on this I sought to estimate their relative contributions. During this work an unexpected, localised mutation pattern was discovered and subsequent analysis demonstrated some loci to be highly susceptible to small segmental deletions in a subset of cancers. In the absence of a justifiable model of neutral segmental deletion it was not possible to infer whether these major mutations could be considered passengers or drivers of cancer progression. In contrast, an advantage of the evolutionary approach applied to nucleotide substitutions in protein coding sequences is that there is a justified model of neutral evolution (synonymous changes). Using this approach, I have not only been able to detect genes harbouring putative cancer driver mutations, but have also found evidence for genes subject to purifying selection in cancers where potentially disruptive mutations appear to be deleterious to cancer progression. Such genes, if they are non-essential in the adult organism, could provide a novel type of target for anti-cancer therapeutics.

Acknowledgements

Firstly, I would like to acknowledge and thank my supervisor Martin Taylor for his guidance and support throughout the course of my PhD. I would also like to thank the members of my thesis committee panel, Malcolm Dunlop, Colin Semple and Andy Sims for their comments and helpful suggestions, and the members of Evogen for all their help and advice. Particular mention to Alison Meynert for her supervision and Bioinformatics advice over the course of my studies, and for teaching me how to Perl. I would like to acknowledge the MRC Capacity Studentship for funding my studies over the last four years.

I would like to acknowledge Alison Meynert for developing and running the TCGA realignment and variant calling pipeline implemented in this project.

Harriet Kemp and Sarah Rennie have been an amazing support-network, especially during thesis writing. I would also like to thank Rob Young and Sara Periconne for their positive encouragement and programming assistance.

Finally thanks to my family and friends for all their support throughout my studies, and to Michelle Rennie for taking the time to read my thesis in its early stages and offer feedback.

Contents

List of Figures

List of Tables

Listings

Abbreviations

BAM	**B**inary **A**lignment **M**ap
BLCA	**Bl**adder Urothelial **Ca**rcinoma
BRCA	**Br**east Invasive **Ca**rcinoma
CARC	**Carc**inoid
CESC	**Ce**rvical **S**quamous **C**ell and Endocervical Adenocarcinoma
CLL	**C**hronic **L**ymphocytic **L**eukemia
COAD	**Co**lon **Ad**enocarcinoma
CRC	**C**olo**r**ectal **C**arcinoma
DLBCL	**D**iffuse **l**arge **B**-**c**ell **l**ymphoma
DNA	**D**eoxyribo**n**ucleic **A**cid
ESO	**Eso**phageal adenocarcinoma
FDR	**F**alse **D**iscovery **R**ate
GBM	**G**lio**b**lastoma **M**ultiforme
HNSC	**H**ead and **N**eck **S**quamous **C**ell Carcinoma
KIRC	**Ki**dney **R**enal **C**lear Cell Carcinoma
KIRP	**Ki**dney **R**enal **P**apillary Cell Carcinoma
LAML	**A**cute **M**yeloid **L**eukemia
LGG	Brain **L**ower **G**rade **G**lioma
LUAD	**Lu**ng **Ad**enocarcinoma
LUSC	**Lu**ng **S**quamous **C**ell Carcinoma
MED	**Med**ulloblastoma
MEL	**Mel**anoma

MM	**M**ultiple **m**yeloma
NB	**N**euro**b**lastoma
OV	**Ov**arian Serous Cystadenocarcinoma
PRAD	**Pr**ostate **Ad**enocarcinoma
READ	**Re**ctum Adenocarcinoma
RHAB	**Rhab**doid tumor
SNV	**S**ingle **N**ucleotide **V**ariant
STAD	**St**omach **Ad**enocarcinoma
TCGA	**T**he **C**ancer **G**enome **A**tlas
THCA	**Th**yroid **Ca**rcinoma
UCEC	**U**terine **C**orpus **E**ndometrial **C**arcinoma

Chapter 1

Introduction

Normally the development of a complex multicellular organism consists of a sequence of tightly regulated and limited cellular divisions, differentiation and programmed cell death (apoptosis). Conversely, in cancer the normal checks and controls on cell division and death are substantially modified [Talavera et al., 2010]. When relatively unrestrained proliferation of cells proceeds, benign growths can result [as reviewed in Stratton et al., 2009]. If those excessively dividing cells acquire the ability to invade adjacent tissue or migrate to distant locations through the blood and lymph systems (metastasise) then a cancer (malignant tumour) has developed [as reviewed in Stratton et al., 2009].

Cancer is a highly diverse disease covering over 100 different diseases originating in most organs of the body [as reviewed in Stratton et al., 2009]. It is widely recognised as the most common genetic disease, with one in three people developing some form of cancer in their lifetime, and for one in five of those affected, the disease is fatal[1]. Cancer is responsible for one in seven deaths worldwide, as the second leading cause of death (following cardiovascular diseases) in high-income countries and third leading cause of death (following cardiovascular diseases and infectious and parasitic diseases) in low-income countries [AmericanCancerSociety, 2015]. Consequently, cancer is an enormous

[1] http://www.cancerresearchuk.org

global health burden, with 14.1 million cases reported worldwide in 2012, 8.2 million of which resulted in death. By 2030 the burden is expected to grow to 21.7 million new cancer cases and 13 million cancer deaths simply due to the growth and ageing of the population [AmericanCancerSociety, 2015]. However cancer survival rates have doubled in the last 40 years. The 10-year cancer survival rate for adults in the UK is currently 50%, with 46% of men and 54% of women adult cancer patients diagnosed in 2010-2011 in England and Wales predicted to survive 10 or more years [Cancer Research UK, 2014]. The cost of cancer in the UK to society as a whole (including costs to NHS and loss of productivity) is estimated to be £18.3 billion a year[2], with the total economic impact worldwide unknown but estimated to be the highest economic loss of all leading causes of death worldwide, totalling hundreds of billions of dollars per year [AmericanCancerSociety, 2015]. This cost is expected to increase due to the increasing cost of cancer therapies as well as the increasing number of new cancer cases. Therefore there are strong medical, social and financial implications to this genetic disease and a better understanding of the mechanisms that lead to cancer is imperative for improving diagnosis, treatment and prevention.

1.1 Cancer is a disease of the genome

Cancer is a disease of the genome, arising as a result of mutation or epimutation occurring in the DNA sequence of the genomes of cancer cells [Strausberg et al., 2001] (Figure 1.1).

Throughout life, the genome within cells of the human body is exposed to exogenous and endogenous mutagens and suffers mistakes in replication, causing DNA damage. In normal cells, these mutations are usually mitigated by DNA repair mechanisms, and DNA damage-induced programmed cell death (apoptosis). However, occasionally one of these mutations alters the function of a critical gene (e.g. DNA repair gene or gene involved in apoptosis), providing a growth advantage to the cell in which it has

[2] http://www.gov.uk

occurred, resulting in the emergence of an expanded clone derived from this cell (Figure 1.1). Acquisition of additional mutations that confer further advantages to the cell, (such as an ability to evade apoptosis, an insensitivity to anti-growth signals, an unlimited replicative potential, sustained angiogenesis and self-sufficiency in growth signals), and consequent waves of clonal expansion, result in the evolution of the mutinous cells that have the capability to invade surrounding tissues and metastasise [as reviewed in Hanahan and Weinberg, 2000, 2011]. As can be seen from Figure 1.1, only a restricted fraction of the cells in a primary tumour are considered to be highly metastatic, with cells in tumour populations usually consisting of different sub-populations of cells with distinct mutations, highlighting how phenotypically and biologically heterogeneous primary tumours are [Yokota, 2000].

Of the mutations that are implicated in cancer development, they can either be inherited germline mutations known as risk alleles that predispose to cancer, or acquired somatic mutations.

1.1.1 Somatic mutations

Mutations that occur during the lifetime of our cells are termed as somatic to distinguish them from the germline mutations that are inherited from parents and passed down to offspring [as reviewed in Stratton et al., 2009]. Cancer requires somatic acquired changes in order to arise. Somatic mutations are thought to occur in the genomes of all normal cells as they proceed through the rounds of cell division that take place during development in utero and in replenishment of body tissues during postnatal life, with additional somatic mutations accumulating during division of cancer cells [as reviewed in Stratton, 2011]. These cancer-causing mutations occur in cancer genes, which can be classified as either recessively acting tumour suppressor genes or dominantly acting oncogenes [as reviewed in Stratton, 2011]. In total there are ∼400 known somatically mutated cancer genes known to contribute to neoplastic change in one or more types of cancer [as reviewed in Stratton, 2011].

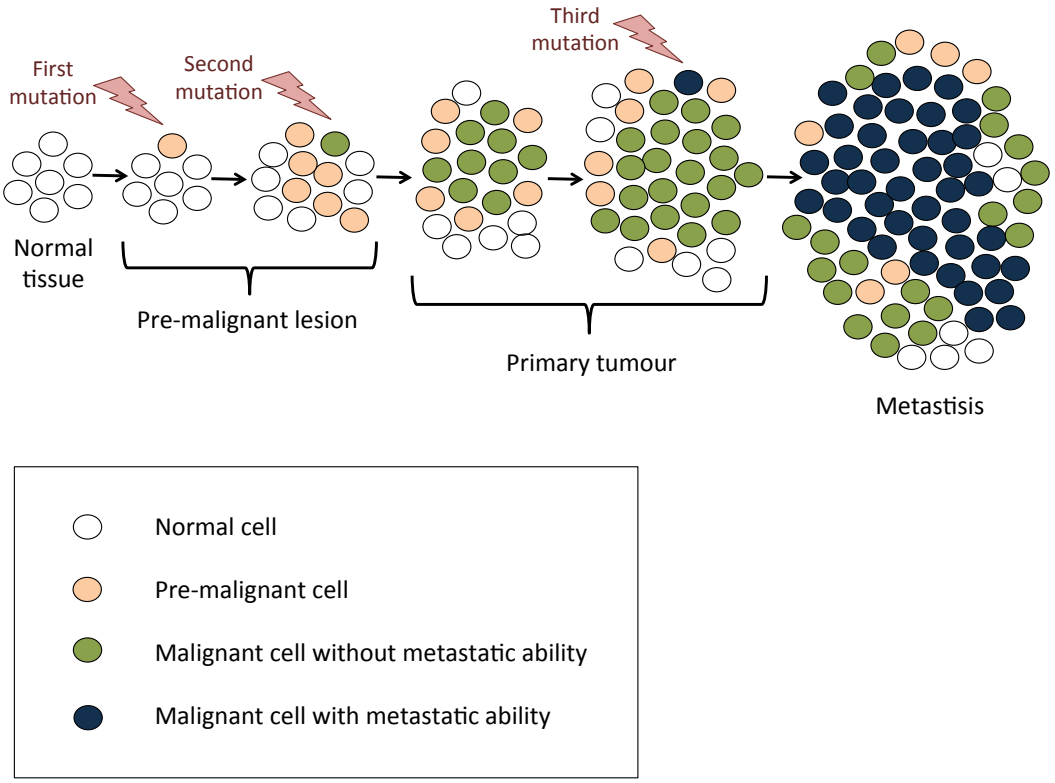


FIGURE 1.1: **Malignant initiation and progression of human cancer.** Pre-malignant (benign) lesions are caused either by genetic alterations or environmental factors which induce clonal expansion of cells. Accumulation of subsequent genetic alterations occur in a single pre-malignant cell, converting the cell into a malignant one which clonally expands and produces a primary tumour. New clones with invasiveness and metastatic ability appear as a result of further acquisition of genetic alterations conferring invasiveness and metastatic potential in cells, to produce fully malignant cells. *Adapted from Yokota [2000].*

1.1.1.1 Tumour suppressors

The normal activity of tumour suppressors is to suppress growth and other behaviour characteristics of cancer, so a loss of that activity is associated with cancer development and progression.

Tumour suppressor genes typically have a recessive nature at the level of the cancer cell, requiring mutation of both parental alleles of the gene, usually resulting in inactivation of the encoded protein in order to lead to loss of function of the gene [Yang et al., 2003].

Recessive cancer genes are characterised by a diverse pattern of mutation types, ranging from single base substitutions to whole gene deletions, with the common outcome of abolishing the protein function [as reviewed in Stratton et al., 2009]. In particular stop codon mutations (stop-gained or stop-lost base substitutions) and frameshift INDELs often result in a loss of gene function, knocking out the function of the encoded protein. Such inactivations can also arise from epigenetic silencing [Vogelstein and Kinzler, 2004].

TP53 is an example of such a tumour suppressor gene, whose normal function as a transcription factor is to upregulate expression of genes involved in cell cycle arrest [Levine, 1997], or to induce apoptosis in cells with damaged DNA independent of transcription [Bálint E and Vousden, 2001]. TP53 is also involved in the regulation of DNA repair, specifically through the nucleotide excision repair of UV damage, and through base excision repair of hydrogen-peroxide induced damage [Smith and Seo, 2002]. Somatic TP53 mutations occur in almost every type of cancer at rates from 38-50% in ovarian, esophageal, colorectal, head and neck, larynx, and lung cancers to about 5% in primary leukemia, sarcoma, testicular cancer, malignant melanoma, and cervical cancer [Olivier et al., 2010]. When the function of TP53 is lost, cells can escape its apoptotic and senescence effects causing a precancerous lesion to become cancerous [as reviewed in Negrini et al., 2010]. Other examples of tumour suppressor genes include RB, whose normal function is to block the cell cycle at the G1 phase [Yang et al., 2003] and BRCA1 which is normally involved in DNA damage response and repair pathways [Yang et al., 2003]. When the normal processes of these genes are knocked out by mutation, the genes take on the characteristics needed to facilitate cancer progression.

Tumour suppressor genes commonly become inactivated in cancer through the deletion of one allele via a gross chromosomal event, such as the loss of an entire chromosome or a chromosome arm, coupled with an intragenic mutation of the other allele [Vogelstein and Kinzler, 2004]. This can result in a loss of heterozygosity (LOH). This refers to pre-tumour cells that are heterozygous for alleles of a tumour suppressor gene (e.g. one normal and one mutant allele), but the tumour cells have lost the normal functional tumour suppressor allele meaning that they are no longer heterozygous. This can occur

either by a mutation that is inherited or a sporadic mutation that inactivates one copy of the gene in a somatic cell. A second mutation, for example the deletion of a large region or loss of entire chromosome in the other copy of the gene would lead to LOH in the tumour suppressor gene and potentially have a proliferative advantage contributing to tumourigenesis. Tumour suppressor genes can therefore be found by looking for LOH in a tumour. If LOH is frequently seen in a specific region of the genome in a particular tumour type, this suggests that a tumour suppressor gene that plays an important role in the development of that tumour may be present in this region. An example of LOH in tumour suppressor genes acting in cancer is in the DNA repair gene BRCA2, which is known to often carry an inactivating recessive germline mutation on one allele of the gene, conferring susceptibility to breast and ovarian cancer. This gene is heterozygous at the germline mutant site, however a subsequent inactivating mutation on the unaffected functioning allele of the gene can result in a LOH removing BRCA2 activity leading to genome instability, a known mutator phenotype that can lead to subsequent mutations driving the development and progression of cancer.

1.1.1.2 Oncogenes

Oncogenes are mutated versions of proto-oncogenes, whose normal role is typically in promoting cell growth in the presence of a relevant growth signal [Yang et al., 2003].

The conversion of a proto-oncogene into an oncogene is often caused by a gain of function mutation, resulting in activation of the encoded protein, conferring enhanced or new activity [as reviewed in Stratton et al., 2009]. Oncogenes have a dominant nature, requiring only one allele of the gene to be mutated in order to contribute to cancer development [as reviewed in Stratton et al., 2009]. Mutations occurring in oncogenes are therefore not likely to be knockout mutations. These changes are likely to be quite specific in order to mediate increased expression of the gene, so high degrees of mutation clustering are often seen in oncogenes. The repertoire of cancer-causing mutations in oncogenes is more constrained compared to that of tumour suppressor genes, in regards to the type of mutation class and the location of the mutation in the gene. For example,

missense amino acid changes, in-frame indels and gene amplifications are all common mechanisms by which oncogenes are activated, with most oncogenes activated through genomic rearrangements such as translocations creating fusion genes [as reviewed in Stratton et al., 2009].

Examples of known oncogenes are the RAS genes (KRAS, HRAS, and NRAS), which encode proteins with guanosine-nucleotide binding activity and intrinsic guanosine triphosphatase activity. When these genes are mutated in codon 12, 13, or 61, they encode a protein that remains in the active state and continuously transduces signals by linking tyrosine kinases to downstream serine and threonine kinases, inducing continuous cell growth. Mutations of KRAS are common in carcinomas of the lung, colon, and pancreas, whereas mutations of NRAS occur principally in acute myelogenous leukemia and the myelodysplastic syndrome [Croce, 2008]. β-catenin is also an oncogene, involved in cell growth and commonly mutated in colorectal cancer [Polakis et al., 1999]. Once activated, β-catenin promotes tumour progression through persistent activation of one of its downstream targets [Polakis, 1999].

Over 80% of known cancer genes are oncogenes. However this estimate is subject to ascertainment bias, since most of these oncogenes are caused by chromosomal translocations and most of the ~400 known cancer genes have been identified through cytogenetic analyses biased towards detecting chromosomal rearrangements such as translocations [as reviewed in Stratton, 2011].

1.1.2 Heritable cancer susceptibility

As well as somatically acquired mutations which are needed for the initiation of cancer, germline mutations can also contribute by providing susceptibility alleles that increase the lifetime risk of developing cancer. Germline mutations can influence cancer susceptibility by directly altering the growth of the cancer clone, altering the mutation rate in somatic cells or modulating the metabolism of carcinogens. These mutations are

present in the fertilised egg from which the individual develops and are therefore found in all somatic cells of the body [as reviewed in Stratton, 2011].

Well-known examples are BRCA1 and BRCA2, both of which have been identified as genes known to contain germline mutations that confer a susceptibility towards developing breast and ovarian cancer [Boulton, 2006, Cancer Genome Atlas Research Network, 2011, Stephens et al., 2009]. BRCA2 was first discovered to be a germline gene containing mutations predisposing to breast cancer in 1995 by Wooster et al. [1995]. Complete loss of either of these genes results in genome instability involving segmental rearrangement, deletion and loss of heterozygosity in cancer [Waddell et al., 2015].

1.1.3 Viral oncogenes

Some viral genomes encode oncogenic proteins (viral oncogenes) and infection with such viruses is associated with specific cancers. Oncogenic viruses can be broadly categorised based on the nature of their genome as DNA or RNA viruses. Viral genomes are commonly integrated into the cancer cell genome in order to become oncogenic [Zheng, 2010].

Examples of DNA viruses known to contribute to cancer are the human papillomavirus which can lead to cervical cancer [Dürst et al., 1983] and head and neck cancer [Talbot and Crawford, 2004], the Epstein Barr virus which is associated with various lymphoid malignancies, human immunodeficiency virus (HIV) linked to non-Hodgkin lymphoma, hepatitis B virus which can lead to liver cancer, human T-cell lymphocytic virus type 1 associated with adult T-cell leukemia/ lymphoma and human herpes virus 8 associated with Kaposi's sarcoma [Talbot and Crawford, 2004]. Hepatitis C is an example of an RNA virus that is associated with liver cancer [Shiffman and Benhamou, 2015].

In addition to infection by exogenous viral genomes, the germline human genome contains many integrated viral genomes (endogenous retroviruses) and other repetitive elements that use the endogenous retroviral replication machinery (including reverse transcriptase) to transpose (insert new copies) into the genome. The activation and

transposition of such endogenous elements has been implicated as a mutational process in some cancers [Lee et al., 2012].

1.1.4 Contagious cancers

Two highly unusual cancers exist that have overcome the limitations of existing within a single host by gaining the ability to spread between individuals [as reviewed in Murchison, 2008]. Tasmanian devil facial tumour disease (DFTD) and canine transmissible venereal tumour (CTVT) are the only known naturally occurring clonally transmissible cancers, caused by direct communication from one host to another [Welsh, 2011].

DFTD is transmitted during fighting and courtship battles spread by the physical transfer of facial tumour cells from one animal to another as an allograft. CTVT has also been shown to be transmitted as an allograft, passed on during mating in which malignant tumour cells from one dog are directly transferred to another dog via coitus, licking, biting and sniffing of affected areas (genitals, nose and mouth) [Welsh, 2011].

DFTD is often fatal within 6 months, whereas most cases of CTVT are eventually cleared by the host dog immune system. Tasmanian devil populations appear to be lacking in genetic diversity which seems to result in the transplanted tumours cells not being rejected, leading to progression of the cancer and ultimately a fatal outcome. Although CTVT is ultimately rejected, it has been shown to develop into metastatic CTVT in immunosuppressed dogs, supporting the idea that this contagious cancer is cleared via immune-mediated rejection in healthy dogs [Welsh, 2011]. Accordingly, tumours can be readily transmitted to immunocompromised organisms such as the nude mouse [Xie et al., 2015]. This highlights the specific challenge that human cancer presents - it is host-derived, so from its very starting point it is difficult for the immune system to discriminate from healthy cells. There is also a similar challenge in the treatment of human cancer; the need to identify a drug or other agent that will specifically target a cancer cell without affecting a healthy cell, when both types are so

closely related. It is for these reasons we need to study the changes in cancer in detail, as it is those specific changes that represent the targets for treatment.

1.2 The functional impact of mutations

Although somatic mutations are required for cancers to arise, not all somatic mutations in a cancer genome contribute to the development of the tumour [as reviewed in Stratton et al., 2009]. Based on this, each somatic mutation in a cancer cell genome can be classified according to its functional impact, termed as either a "driver" or "passenger" mutation.

The main focus of cancer genomics is in identifying the functionally important driver mutations that are important in driving cancer development, and distinguishing them from the passengers that are inconsequential to the cancer. This is vital in order to understand the mechanisms that lead to cancer development and progression.

1.2.1 Driver mutations

Driver mutations are causally implicated in the development and progression of oncogenesis [Pleasance et al., 2010a], occurring in cancer genes such as tumour suppressor genes or oncogenes [Greenman et al., 2007]. It is these driver mutations that are responsible for removing function from a tumour suppressor gene or adding function to a proto-oncogene (possibly by dis-inhibiting the encoded protein's function through activation).

Genes containing driver mutations are assumed to be more recurrently hit by mutation in cancer than other genes. TP53 is an example of a validated cancer gene known to be hit recurrently by driver mutations in many cancers [Forbes et al., 2011].

The number of driver mutations and hence the number of mutated cancer genes required for an individual cancer to develop is speculated to be around five. However, this

number is thought to be lower for hematopoietic cancers [as reviewed in Stratton, 2011].

1.2.2 Passenger mutations

Many other somatic mutations also accumulate in tumours, however they are not implicated in oncogenesis as they do not confer a growth advantage to the cancer, and are therefore termed passenger mutations. The passenger mutations present in cancer cells are much more abundant than driver mutations [Bignell et al., 2010]. Although passenger mutations do not help or hinder the cancer they can occur in genes that also harbour driver mutations.

Although the primary target of cancer studies is to identify the driver mutations that play a role in cancer initiation and progression with passenger mutations viewed as a complication in the search for causal mutations, these passenger mutations can also provide a rich source of information reflecting the specific mutational processes (rather than selection) underlying cancer in somatic cells, and hence are an important by-product [Rubin and Green, 2009]. Such mutational trends could reflect passenger mutations arising before carcinogenesis that are carried along by subsequent clonal expansion as well as passenger mutations that occur within cancer cells [Rubin and Green, 2009].

1.3 Types of mutation event

Somatic driver mutations can exist in the form of point mutations (single nucleotide variants (SNVs)), small insertions and deletions (indels), copy number variations or large-scale genomic rearrangements caused by breakage and abnormal rejoining of DNA [Cancer Genome Atlas Research Network, 2008, Strausberg et al., 2001]. These mutations can occur in either the coding regions or non-coding regions of the human genome.

1.3.1 Single nucleotide variants (SNVs)

Single nucleotide variants (SNVs) are the most common class of mutation in cancer [as reviewed in Meyerson et al., 2010]. This type of mutation refers to a point substitution in which a single nucleotide in the genome sequence is mutated. On average SNVs occur at a rate of one per million nucleotides in human malignancies, however this number varies between tumour types and even between individuals within tumour types [DeVita et al., 2010].

Point substitutions can be further sub-categorised into coding and non-coding mutations. Non-synonymous and synonymous mutations are both terms used to describe mutations appearing in coding regions of the genome. Non-synonymous changes in the DNA are those which cause a change to the encoded protein, and these are generally known as missense mutations. Synonymous mutations, although present in coding regions of the genome, do not affect the encoded protein, due to the redundancy seen in the genome with more than one codon coding for the same amino acid. Synonymous mutations are generally assumed to not be as functionally important as non-synonymous mutations are hence selectively silent, due to the fact that they do not change the encoded protein [Massingham and Goldman, 2005]. All synonymous mutations are therefore considered to be passengers with no role in oncogenesis [Youn and Simon, 2011]. However it has been recently shown that synonymous mutations are not necessarily selectively silent, and may be advantageous to cancer cells. Supek et al. [2014] found that dosage-sensitive oncogenes have mutations in their 3' UTRs that are under selection.

Along with missense mutations, stop-gained (nonsense) and stop-lost mutations are also considered non-synonymous single nucleotide mutations. A stop-gained mutation refers to a point mutation which has resulted in a premature stop codon, and will likely produce a truncated protein. A stop-lost mutation refers to the loss of a stop codon, which can lead to reduced or a loss of protein function.

Substitutions can be categorised into transitions and transversions depending on the nucleotide change occurring, and more specifically on the structure of the base component nucleotide (since nucleotides consist of a phosphate group, a pentose sugar and an amino base) (Figure 1.2). Transitions involve a substitution between two purines or two pyrimidine bases, whereas transversion involve the change of purine for a pyrimidine or vice versa. Purine bases consist of two carbon rings and are larger than transversions which have only one carbon ring.

These are twice as many possible transversions that could potentially occur in the genome, however transition mutations occur at a significantly higher rate than transversion mutations in normal human genomes. This difference is attributed to the molecular mechanisms by which transitions are generated including the increased rate of C→T transitions at CpG dinucleotides via a process called CpG deamination. This mechanism involves a cytosine methylated at the 5-carbon undergoing hydrolytic deamination to thymine at a relatively high rate. This mutation pattern is also observed in cancer, with transitions found to be more frequent than transversions and CpG dinucleotides shown to be a mutation hotspot, particularly in colorectal cancer [Rubin and Green, 2009].

1.3.2 Micro insertions and deletions (INDELs)

Small insertions and deletions (collectively termed indels) are not as abundant as single nucleotide variants (SNVs), approximately ten-fold less common than single nucleotide mutations [as reviewed in Meyerson et al., 2010], but can have much more damaging effects. They frequently result in severe genetic diseases such as Tay-Sachs disease.

Indels are capable of causing frameshifts if the length of the indel is not divisible by three, by altering the gene's reading frame. Frameshift indels are presumed to be deleterious to gene function [Hu and Ng, 2013], as they result in a completely different translation of the protein, often affecting the normal stop codon which can make the protein abnormally long or short.

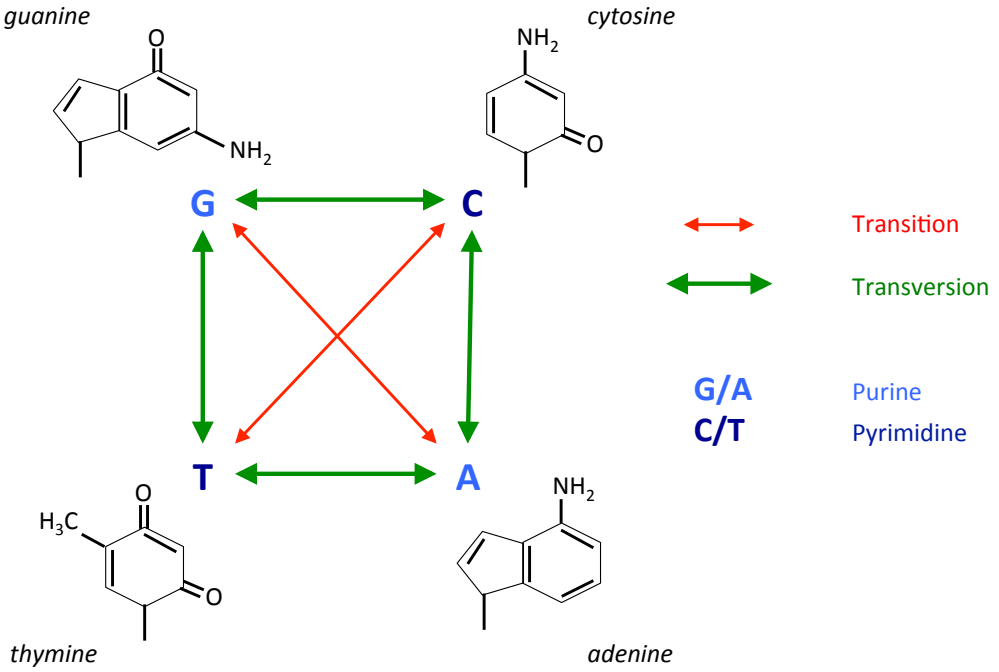


FIGURE 1.2: **Transitions and transversions.** Point mutations can be classified according to their type of amino base substitution: transversions (TR) which involve interchanges of purine for pyrimidine bases, and transitions (TS) which involve interchanges of two-ring purines (A↔G) or of one-ring pyrimidines (C↔T). A = adenine, C = cytosine, G = guanine and T = thymine.

The other type of indel occurring in coding regions are those that have a length that is divisible by three. These are termed "3n indels", and result in insertions or deletions of amino acids in the encoded protein.

1.3.3 Segmental copy number change

Copy number variants (CNVs) are a form of structural variation in the genome that result in the cell possessing an abnormal number of copies of one or more sections of DNA. Copy number variants can range from the gain or loss of chromosome arms (tens of megabases) to focal amplifications and deletions (tens of kilobases) Beroukhim et al. [2007, 2010], Bignell et al. [2010, 2004], Mullighan et al. [2007], Weir et al. [2007], Zhao et al. [2004, 2005].

In cancer, amplifications have been discovered in oncogenes such as MYC in aggressive forms of breast cancer, correlating with poor prognosis and distant metastases [Singhi et al., 2012]. Deletions in tumour suppressor genes such as PTEN in prostate cancer have also been revealed [Dong, 2001]. A survey for somatic copy-number alterations in cancer in Beroukhim et al. [2010] revealed repeatedly deleted and amplified sites in the BCL2 family of apoptosis regulators and the NF- pathway.

1.3.4 Translocation

Chromosomal translocations are one of the most common types of genetic rearrangement [Nambiar and Raghavan, 2011]. A translocation is an abnormal chromosome region containing rearranged genetic material, usually from two non-homologous chromosomes [as reviewed in Bunting and Nussenzweig, 2013]. Some translocations involve the joining of a break to a sequence in the same chromosome (intra-chromosomal), however it is inter-chromosomal translocations that are usually recurrently seen in cancer cells.

A minority of translocations form gene fusion events when two previously separated genes become fused [as reviewed in Bunting and Nussenzweig, 2013], and most often this will cause a loss-of-function phenotype due to the genes being broken by the translocation, sometimes inactivating tumour suppressors. For example the gene fusion TEL1-AML can repress the expression of TEL1, a tumor suppressor gene [Nambiar and Raghavan, 2011]. However, gene fusions can also result in oncogenes in which an activated form of the protein is coded for by the fusion gene [Nambiar and Raghavan, 2011].

A chromosome that contains a translocation is termed a ‘derivative chromosome’, and most are known to contain balanced reciprocal translocations. Reciprocal translocations describe the exchange of genetic material between two chromosome arms. These can be classified as either balanced or unbalanced, depending on whether the copy number is affected. A balanced translocation has no affect on the overall copy number [as

reviewed in Bunting and Nussenzweig, 2013]. Another type of translocation is the Robertsonian translocation in which the long arms of two acrocentric chromosomes are joined around a single centromeric region [as reviewed in Bunting and Nussenzweig, 2013]. Acrocentric chromsomes are those with centromeres situated near one end of the chromosome, resulting in one chromosome arm being much shorter than the other. There are five pairs of acrocentric chromosome pairs present in humans: 13, 14, 15, 21, 22.

The non-homologous end joining (NHEJ) DNA repair pathway has been shown to contribute to the appearance of somatic chromosomal translocations by increasing the amount of genomic instability, particularly when the much less error-prone homologous recombination DNA repair pathway is defective [as reviewed in Bunting and Nussenzweig, 2013]. Classical non-homologous end-joining (C-NHEJ) causes a low rate of translocations, however in its absence a NHEJ pathway called alternative end-joining (A-EJ) or microhomology-mediated end-joining (MMEJ) becomes active and produces an increased number of chromosome translocations. It has also been shown that microhomology-based mechanisms are responsible for a minority of *de novo* human translocations [as reviewed in Bunting and Nussenzweig, 2013]. *De novo* mutations are those that are present in every cell of the body, much like a germline mutation, however they are not present in the parents as a germline mutation would be expected to be. This is due the mutation having occurred in a germ cell of one of the parents, or in the fertilised egg. Supporting this mechanism of NHEJ as a source of translocation, it has been shown that translocations between two sites are highly dependent on the frequency of double-strand breaks, which are known to be repaired by NHEJ. For example, when p53, which promotes apoptosis in cells with double-strand breaks, is deleted, the overall frequency of translocation increases [as reviewed in Bunting and Nussenzweig, 2013], suggesting that an increase in double-strand breaks contributes to an increased rate of translocation. Active transcription also correlates with translocation, with more transcribed genes prone to translocation than transcriptionally silent genes [as reviewed in Bunting and Nussenzweig, 2013].

Chromosomal translocations, and their corresponding gene fusions, have an important role in the initial steps of tumorigenesis [as reviewed in Mitelman et al., 2007]. Historically most gene fusions were found in haematological malignancies, however many have since been found in solid tumours such as prostate cancer [as reviewed in Bunting and Nussenzweig, 2013]. The first translocation identified in human neoplasia was t(9;22)(q34;q11), signifying a translocation between chromosomes 9 and 22 with breakpoints in bands 9q34 of BCR and 22q11 in ABL1 respectively, resulting in a fusion gene, found in almost all cases of chronic myeloid leukemia (CML) [as reviewed in Mitelman et al., 2007]. This translocation is known as the Philadelphia chromosome and is an example of a balanced, reciprocal translocation between two non-homologous chromosomes, which causes overexpression of ABL1 [as reviewed in Mitelman et al., 2007]. This recurrent rearrangement was detected using cytogenetics, which before the advent of second generation sequencing was used to identify chromosomal rearrangements in relatively simple genomes such as those of leukaemias, lymphomas and sarcomas [as reviewed in Meyerson et al., 2010]. In epithelial cancers the first two major recurrent translocations identified were the TMPRSS2-ERG translocation in prostate cancer and the EML4-ALK translocation in non-small cell lung cancer [as reviewed in Meyerson et al., 2010].

The tyrosine kinase activity of BCR-ABL1 essential for cellular transformation has been successfully targeted therapeutically in an example of personalised medicine. The product of this translocation, a protein known as bcr-abl-tyrosine kinase, can be inhibited by a specific drug called imatinib mesylate [Koss, 2007]. It is the tyrosine kinase activity of BCR-ABL1 which is essential for cellular transformation [Deininger et al., 2005], and that is also inhibited by the protein kinase inhibitor imatinib.

Translocations are often associated with the phenomenon known as chromothrypsis in which one or a few chromosomes in a cancer cell harbour tens to hundreds of clustered genome rearrangements [as reviewed in Forment et al., 2012]. This complex event is likely to arise through chromosome breakage and inaccurate reassembly [as reviewed in Forment et al., 2012]. The mechanism was first observed in chronic lymphocytic

leukaemia (CLL), in which 42 genomic rearrangements were discovered in the long arm of chromosome 4 through paired-end sequencing of a CLL genome [as reviewed in Forment et al., 2012]. Chromothripsis occurs commonly in various human cancers, including melanomas, sarcomas, gliomas, colorectal, oesophageal, renal and thyroid cancers [as reviewed in Forment et al., 2012].

Chromothrypsis can also involve other chromosomal rearrangements as well as translocation, such as tandem duplications, interstitial deletions and chromosomal inversions [Zhang et al., 2013]. Similarly fusion genes are not just caused by translocations but can also be the result of different types of chromosomal rearrangement.

1.3.5 Epimutation

While genomics involves the study of heritable or acquired alterations in the DNA sequence, epigenetics is the study of heritable changes in gene activity caused by mechanisms other than DNA sequence changes. Mechanisms of epigenetic activity include DNA methylation, small RNA-mediated regulation, DNA-protein interactions and histone modification.

Over the lineage of mitotic cell divisions from the progenitor fertilised egg, the cancer genome can acquire epigenetic mutations. These mutations alter chromatin structure and gene expression, manifesting at the DNA sequence level by changes in the methylation status of some cytosine residues [as reviewed in Stratton et al., 2009]. Epimutation can also cause the abnormal transcriptional repression of active genes or the abnormal activation of usually repressed genes in cancer.

Epigenetic mutations often affect tumour suppressor genes by functionally mirroring loss of function mutations. For example, BRCA1 function is inactivated through BRCA1 promoter methylation, causing transcriptional silencing, in a substantial number of triple-negative breast cancers [Xu et al., 2013]. Many cancer genes uncovered by recent systematic sequencing encode proteins involved in chromatin modification and remodelling. For example, the tumour suppressor genes SETD2 and MLL2 and

the oncogene EZH2 are histone H3 methylases. The tumour suppressor genes KDM6A and KDM5C are histone H3 demethylases. The tumour suppressors ARID1A, PBRM1, BAP1, ATRX and DAXX are components of protein complexes responsible for restructuring chromatin, and DNMT3A is involved in the maintenance of cytosine methylation [as reviewed in Stratton, 2011]. These findings support a potentially important link between somatic mutation and epigenetic changes in some cancers.

1.4 Sources of mutation

Somatic mutations occur in the genomes of all dividing cells, both neoplastic and normal, arising as a result of misincorporation during DNA replication or through exposure to endogenous or exogenous mutagens [Greenman et al., 2007].

In normal cells, these mutations either confer no growth or survival advantages to the cell or they are repaired by endogenous DNA repair mechanisms. If the damaged DNA is not able to be repaired, then the cell is programmed for apoptosis. However, in cancer cells, mutations in genes that control DNA repair or apoptosis commonly occur, meaning that other cancer-causing mutations in oncogenes and tumour suppressor genes can accumulate in the cell, giving rise to a cancer.

1.4.1 Mutator phenotype

Many mutations are required for the neoplastic transformation of cells, therefore cancers are often associated with an elevated mutation rate compared to normal cells; this is known as a "mutator phenotype" [Loeb, 2001]. A mutator phenotype refers to a higher mutation rate than normal during the part of the cell lineage in which predecessors of the cancer cell already show phenotypic evidence of neoplastic change [as reviewed in Stratton et al., 2009]. The rate of mutation acquisition can be increased by exogenous and endogenous exposures that cause DNA damage, and can be mitigated by DNA

repair processes, so in the event that repair processes fail, the somatic mutation rate may also increase [as reviewed in Stratton, 2011].

There is also great variation in mutation rates between different tumour types, with some tumour types exhibiting much higher mutation rates than others. There are between 1000 and 10,000 somatic mutations in the genomes of most adult cancers, such as breast, ovarian, colorectal, pancreas and glioma. However, cancers such as melanoma and lung cancer have been shown to carry more than 100,000 mutations, with some other cancers such as acute leukemias, medulloblastomas, carcinoids and testicular germ cell tumours having relatively fewer mutations [as reviewed in Stratton, 2011]. Cancers with lower mutation rates tend to be childhood and young adult cancers, so the low mutation frequency could be explained by the fact that the neoplastic cell has been through relatively few DNA replications. For example, haematopoietic cancers carry less than one mutation per million bases [as reviewed in Meyerson et al., 2010]. However it is also possible that a mutator phenotype is not necessary in these cancer types for neoplastic transformation to occur. The very high prevalence of mutation rates in some cancers is likely to be attributable to heavy mutagenic exposures such as UV light and tobacco, the presence of known defective DNA repair mechanisms and therapy with DNA-damaging agents [as reviewed in Stratton, 2011, Stratton et al., 2009]. For example, melanomas known to be induced by ultraviolet radiation possess on average ten mutations per million bases [as reviewed in Meyerson et al., 2010]. Additionally, cancers originating from epithelial cells such as colorectal, lung and gastric cancers are also seen to exhibit a high mutation prevalence, thought to be due to the high turnover of these cells and the fact that they are subject to recurrent exogenous mutation exposure [Greenman et al., 2007]. These examples display the ability of environmental exposures to increase the mutator phenotype further so that cells acquire mutations much more quickly, increasing the lifetime risk of that cancer.

However, the variation in mutation rate between different cancers is not only attributable to the mutation rate at the cell divisions that have taken place between

the fertilised egg and the cancer cell; the number of mitoses in the lineage also account for the differences in mutation prevalence [as reviewed in Stratton, 2011].

Cancer studies have also shown that there is wide variability in the number of somatic changes within classes of cancer as well as between different cancer types [Greenman et al., 2007], reflecting the different mutational processes operative in individual cancers within a specific tumour type. For example tumours with mutations in mismatch repair genes have higher mutation rates [as reviewed in Watson et al., 2013].

1.4.1.1 Mutational spectra

The mutational spectra are the mutational patterns incorporating information such as the numbers of each class of mutation, the DNA sequences around each mutated base (the sequence context) and in transcribed regions whether the transcribed or untranscribed strand is preferentially mutated [as reviewed in Stratton, 2011]. The mutational spectra therefore refers to the specific type of mutator phenotype.

For example, a specific mutator phenotype is observed in colorectal and endometrial cancers, which typically exhibit elevated rates of acquisition of single nucleotide mutation rates and indels at polynucleotide tracts; this is due to mutations in MLH1 and MSH2 which cause defective DNA mismatch repair and subsequently microsatellite instability [Lengauer et al., 1998]. A specific mutational signature is also known to be associated with chemotherapy treatment. For example, in recurrent gliomas previously treated with the DNA alkylating agent temozolomide, a very high number of mutations with a signature typical of the therapeutic agent is observed [as reviewed in Stratton, 2011, Stratton et al., 2009].

Specific mutational spectra are also observed in cancers known to be affected by environmental mutagenic exposures. For example, lung cancers exhibit many C:G>A:T transversions, which is a pattern similar to that induced by tobacco carcinogens in experimental systems, suggesting that this particular mutator phenotype is caused by the environmental carcinogen. Similarly, hepatocellular cancers also show C:G>A:T

mutations that are likely to be caused by aflotoxins, which are known causative agents in liver cancer development. Skin cancers show mostly C:G>T:A mutations which is a known mutation pattern resulting from UV light, which again is a known aetiological agent in skin cancer [as reviewed in Stratton, 2011].

As well as environmental exposures increasing mutation rate in cancer genomes, the sequence context can also play a role, with regions of the genome known as "hotspots" shown to be more susceptible to spontaneous mutations, showing that sporadic mutations do not occur randomly across the genome. A common source of this elevated mutation rate in hotspots is CpG deamination, the most frequent mutation observed in the human genome. This is the process of a methylated cytosine being mutated to thymine at CpG sites (Figure 1.3). For example, gastrointestinal tumours such as oesophageal, colorectal and gastric have a high frequency of transitions at CpG dinucleotides, which may be caused by CpG deamination reflecting increased methylation levels in these tumours [as reviewed in Watson et al., 2013]. An additional example of cancer types with a known mutational signature of a carcinogenesis mechanism dependent on sequence context is observed in cervical, bladder, some head and neck cancers. These cancer types frequently exhibit mutations at cytosines in the context of TpC dinucleotides, changing the C to either a T or a G or (less often) an A. This signature is characteristic of mutations caused by the APOBEC family of cytodine deaminases [Lawrence et al., 2013].

Some cancers exhibit a mutator phenotype showing very large numbers of mutations with specific mutational signatures seen in only a subclass of tumour types for which the underlying pathogenesis is not understood. For example, in a subset of breast cancers, sequencing of the coding sequence of 518 kinase genes revealed C>G mutations occurring almost exclusively at cytosines following a thymine (TpC dinucleotides, with T immediately 5' to C) [Stephens et al., 2005]. This signature has also been seen in other cancer types such as lung and ovarian cancer suggesting that the mutational process underlying this profile is not specific to a single tumour type, and could be due to a DNA repair defect or a shared mutagenic exposure [Greenman et al., 2007]. Other

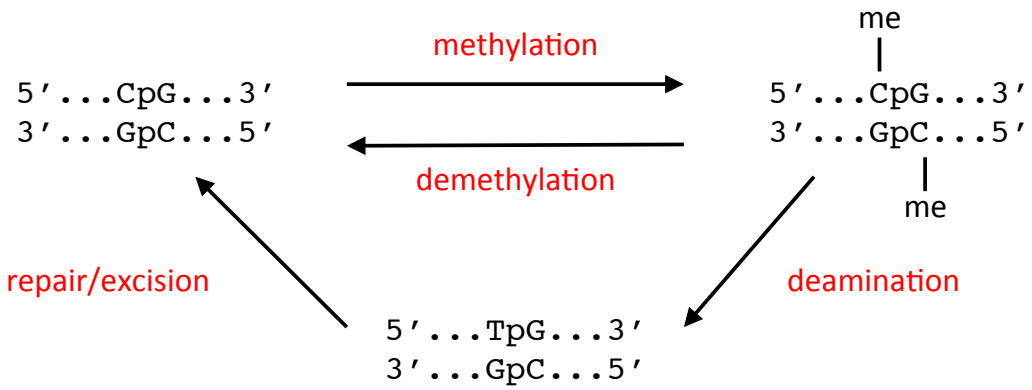


FIGURE 1.3: **CpG deamination.** A methylated cytosine in a CpG island is mutated to a thymine during the process of CpG deamination.

less well characterised mutational spectra can also result in chromosomal abnormalities affecting chromosome number or increased rates of genomic rearrangement [Lengauer et al., 1998]. For example, some breast cancers exhibit frequent tandem duplications, for which the underlying mutagenic or DNA repair processes are unknown [as reviewed in Stratton, 2011].

1.4.2 DNA replication errors

In normal cells, DNA is replicated with extraordinary fidelity ($\sim 10^{-10}$ mutations per base pair per cell division), achieved through a combination of polymerase base selectivity, intrinsic proofreading 3'→5' exonucleolytic activity and base-base mismatch correction. There are three primary eukaryotic DNA polymerase replicative enzymes functioning at DNA replication forks: Pol α (primase), Pol δ and Pol ϵ. Pol δ and Pol ϵ are responsible for DNA synthesis on the lagging and leading strand respectively, with polymerase domains (POL) within these polymerases discriminating between correct and incorrect dNTPs prior to phosphodiester bond formation. DNA polymerase errors occurring during DNA synthesis are corrected by the proofreading exonuclease (EXO) present in each polymerase. Any base-base mispairs or primer-template slippage errors

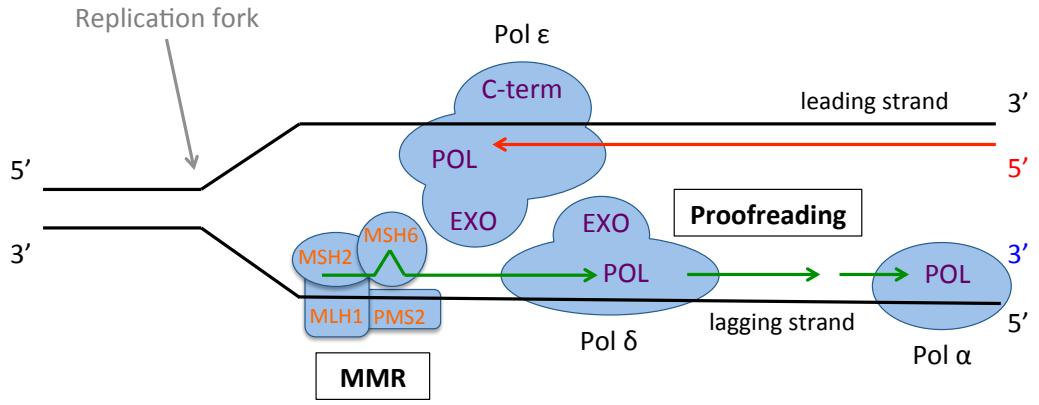


FIGURE 1.4: **DNA replication.** DNA is synthesised by polymerases α, δ and ϵ. The polymerase domains (POL) of these three polymerases recruit dNTPs to the growing DNA strand. The 3'→5' proofreading exonucleases (EXO) present in Pol δ and Pol ϵ correct any errors made by the polymerase domains. MMR corrects base-base mispairs not fixed by the EXO domains, acting on both the leading and lagging strands (shown only on the lagging strand here). *Adapted from [Preston et al., 2010].*

not successfully corrected by the proofreading machinery are repaired by MMR genes by excising the errant bases (Figure 1.4). The role of Pol α is in priming both leading and lagging strand syntheses, copying relatively short stretches of DNA, whereas Pol δ and Pol ϵ synthesise the bulk of chromosomal DNA during cell division [Preston et al., 2010].

In cancer it is the rare somatic copying errors made through DNA replication that can go on to contribute to the initiation and development of cancer. If these mutations are in the DNA replication machinery, such as polymerase genes, this can cause more mutations to occur during DNA replication which in turn increases the risk of developing cancer. The more mutations in different genes and the more that goes wrong in the cell, the higher the chance of a cancer developing.

1.4.2.1 Base selectivity in Pol δ and Pol ϵ

It is estimated that each time a diploid mammalian cell replicates, replicative eukaryotic DNA polymerases make approximately 100,000-1,000,000 errors per day [Preston et al.,

2010], in which a non-complementary base is incorporated into the daughter strand. The majority of these errors are base-base mispairs and ±1 slippage events. These errors must be corrected by proofreading and MMR with almost 100% efficiency to achieve a spontaneous mutation rate of $\sim 10^{10}$ per base pair per cell division, in order to suppress a mutator phenotype that could lead to spontaneous tumorigenesis. Indeed, mutations affecting base selectivity in Pol δ and Pol ϵ have been shown to confer mutator phenotypes in yeast and mice [Preston et al., 2010].

1.4.2.2 Polymerase proofreading

Spontaneous errors made by the leading and lagging strand polymerases, Pol ϵ and Pol δ respectively, trigger proofreading by their intrinsic 3'→5' exonucleases. A third eukaryotic polymerase, Pol γ, also has intrinsic 3'→5' exonucleolytic proofreading activity, however Pol γ is mitochondrial while Pol δ and Pol ϵ are nuclear [Preston et al., 2010]. In yeast, point mutations that selectively inactivate exonucleolytic proofreading activities of Pol δ and Pol ϵ confer moderate to strong mutator phenotypes [Preston et al., 2010]. Mice studies show that a knock-in mutation in the exonuclease domain of Pol δ leads to spontaneous formation of epithelial cancers in lung, thymus and skin; and in engineered mice with defective Pol ϵ proofreading, mice exhibit a susceptibility to developing spontaneous intestinal adenocarcinomas [Preston et al., 2010].

1.4.2.3 Mismatch repair (MMR)

Errors escaping proofreading activity during replication are corrected by MMR machinery. There are two partially redundant pathways of the MMR system, depending on the type of error: mispair mutations are repaired by the MutSα complex and primer-template errors are repaired by either the MutSα or MutSβ complex (Figure 1.5).

The failure to repair the errors escaping proofreading through MMR has been shown to contribute to cancer progression. For example, colorectal, gastric, ovarian, emdometrial and lung cancers often display inactivation of the MMR pathway through mutation of

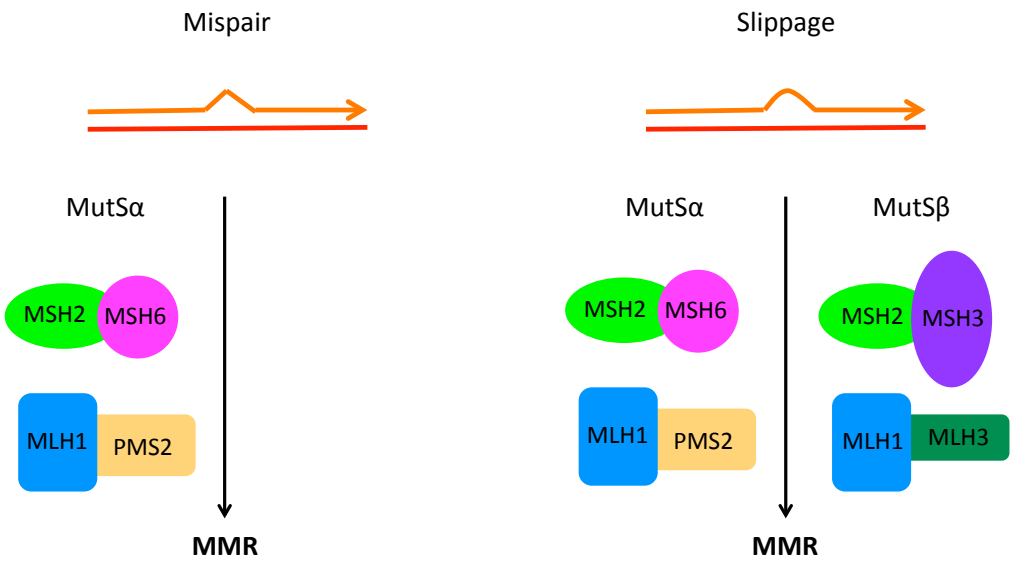


FIGURE 1.5: **MMR pathways.** Errors that escape 3'→5' exonucleolytic proofreading can be corrected by two MMR pathways: MutSα or MutSβ. If the error is a base-base mispair, MutSα is employed to repair, however if a primer-template slippage has occurred then MutSα or Mutβ can be recruited to fix the error. *Adapted from [Preston et al., 2010].*

MMR genes which in turn results in a high incidence of indels at simple sequence repeats known as microsatellite instability (MSI), as well as an increased SNV load [Preston et al., 2010, Supek and Lehner, 2015]. Germline mutations have also been found in MMR genes. For example, mutations in MSH2, the gene encoding one of the key proteins required for MMR in both MutSα and MutSβ are frequently seen in hereditary non-polyposis colorectal cancer (HNPCC) [Preston et al., 2010]. MLH1, another gene required for MMR in both complexes, is also found to have mutations in HNPCC families [Preston et al., 2010]. The majority of HNPCC patients inherit a mutation in either MSH2 or MLH1, and a smaller percentage inherit mutations in PMS2 or MSH6. HNPCC is known to increase the risk of developing colorectal cancer, which develops when the wild-type allele is lost usually through LOH or gene silencing. However, inherited MMR deficiency is only responsible for a small percentage of colorectal cancer cases (1-5%) with most colorectal cancer cases with MSI resulting from acquired defects in MMR, usually due to MLH1 promoter hypermethylation.

MMR deficiency in some cancers has been shown to cause an unusual regional mutation rate variation pattern [Supek and Lehner, 2015], which sheds light on the mechanism of MMR and how it is related to chromatin structure. Schuster-Böckler and Lehner [2012] show that the variation in regional mutation rate observed across the cancer genome is determined by chromatin organisation, with cancer SNV density highly correlating with genetic and epigenetic features of somatic cell chromatin structure. SNV density was found to anti-correlate with early replication timing, associated with open chromatin (euchromatin), and positively correlate with repressive histone modifications, such as H3K9me3 marks associated with closed chromatin (heterochromatin), with mutation rates elevated in more late-replicating heterochromatin-like domains and repressed in early-replicating euchromatin regions [Schuster-Böckler and Lehner, 2012]. The association between chromatin structure and mutation rate variation is also shown to be upheld over diverse tissue types (eg. melanoma, lung cancer, prostate cancer, leukemia), mutation types (eg. transitions vs transversions, CpG vs non-CpG mutations) and genomic regions (eg. genic vs non-genic regions) [Schuster-Böckler and Lehner, 2012]. Supek and Lehner [2015] have used TCGA cancer data to investigate the mechanism responsible for this chromatin-dependent variation in mutation rate across the genome. They concluded that differential DNA repair rather than differential mutation supply is the primary cause of fluctuating mutation rate across the genome. More specifically they have shown that MMR is the underlying factor affecting the varying mutation rate, with MMR-deficient cancers showing no reduction in mutation rates in early-replicating euchromatin compared to late-replicating heterochromatin as is seen in MMR-proficient cancers, with a very high mutation prevalence observed genome-wide in MMR-deficient cancers (Figure 1.6). In MMR-proficient cancers, MMR is therefore more effective at repairing replication errors in euchromatin early replicating regions than in late replicating heterochromatin. This increased efficiency in early replicating euchromatin could be due to increased DNA accessibility to repair machinery in open chromatin, more time available for repair, or the fact that most genes performing essential functions are in euchromatin and replicated early so enhanced MMR has been evolutionarily selected for in these regions of the genome as it is beneficial to the survival of the organism

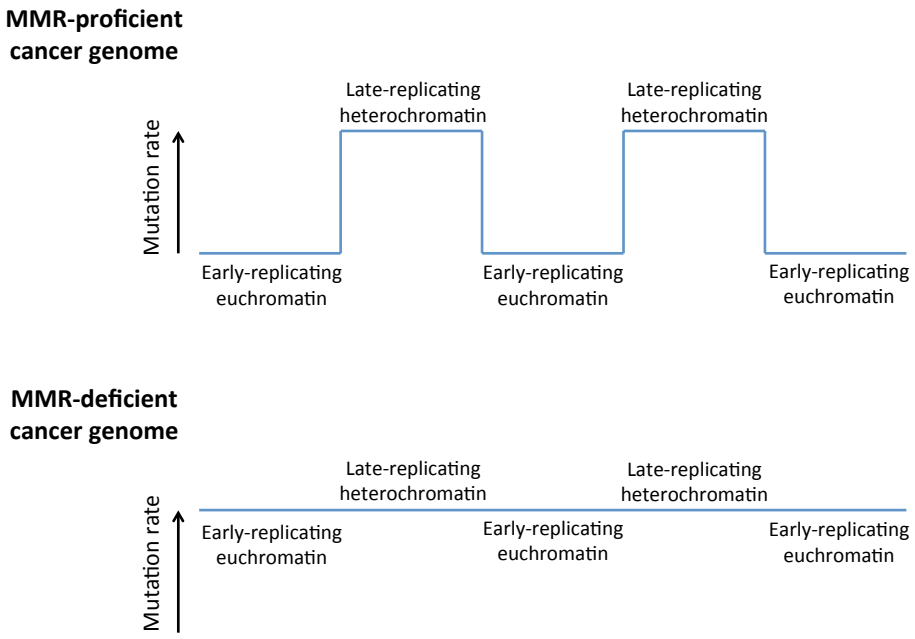


FIGURE 1.6: **Variation in regional mutation rate explained by mismatch repair (MMR).** In MMR-proficient cancer genomes a variable mutation rate pattern is observed in which the mutation rate is increased in late-replicating heterochromatin compared to early-replicating euchromatin. In MMR-deficient cancer genomes, the mutation rate pattern flattens, with an elevated mutation rate observed genome-wide, suggesting that MMR is more effective in early-replicating euchromatin, suppressing the accumulation of mutations in these regions.

[Supek and Lehner, 2015].

1.4.2.4 Proofreading and MMR acting in synergy

Proofreading and MMR have been shown to exhibit a synergistic relationship in the repair of replication errors, with deficiencies in one being compensated by the function of the other, and the loss of both pathways resulting in the accumulation of replication errors [Preston et al., 2010].

1.4.3 Environmental mutagenesis

Somatic spontaneous mutations occur in the genomes of normal cells as they go through the cell cycle at a low frequency (i.e. replication errors). However, the rate at which these cells acquire mutations can be increased by various environmental factors, all of which are known to elevate the lifetime risk of cancer [as reviewed in Stratton, 2011].

The environmental exposure can sometimes induce a specific mutational spectrum in the cancer sample. For example, lung tumours have a high proportion of G→T transversions, attributable to exposure to polycyclic aromatic hydrocarbons from tobacco smoke [as reviewed in Watson et al., 2013]. Melanomas are also known to exhibit a specific mutational spectrum of an elevated proportion of C→T transitions in dipyrimidines, caused by ultraviolet (UV) radiation-induced DNA damage and misrepair [Pleasance et al., 2010a].

1.4.4 Oxidative DNA damage

Reactive oxygen species (ROS) include free radicals and non-radical species, which are produced as by-products of normal cell metabolism, and have the potential to damage proteins, lipids, carbohydrates, and nucleic acids. In normal cells, ROS-induced damage is counteracted by defences including enzymes (catalase, superoxide dismutase, glutathione peroxidase), as well as small molecular weight antioxidants both of endogenous synthesis (glutathione, ubiquinol) and of dietary origin (vitamin C, vitamin E, carotenes) [de Cavanagh et al., 2002]. As a result, a balance is established between the generation and destruction of ROS, minimizing the damage to physiologically relevant biomolecules.

Oxidative stress is an endogenous source of somatic mutation in cancer, and is defined as a disturbance in the balance between the production of ROS and antioxidant defences [Betteridge, 2000], in favour of a higher level of ROS [de Cavanagh et al., 2002]. For example, a mitochondrial antioxidant enzyme MnSOD is found to be present at altered

levels in most types of primary cancers and cell line cancers compared to the control [Oberley, 2002].

1.4.5 Defects in repair processes

Cancer is a disease caused by damage to DNA, therefore how the damaged DNA is repaired plays an important role in the development of cancer. DNA damaged by intrinsic mechanisms such as replication defects and oxidative DNA damage as well as exogenous environmental mutagenesis, are all repaired by various different endogenous DNA damage repair mechanisms depending on the type of mutation incurred [Boulton, 2010]. In cancer, mutations often occur in DNA repair genes, preventing the repair of mutation and increasing the risk of developing cancer.

There are several hereditary human diseases known to predispose to cancer that are caused by defects in DNA repair mechanisms. For example, Xeroderma pigmentosum and hereditary non-polyposis colorectal cancer that confers susceptibility to skin and colorectal cancer respectively [Cleaver, 2005, Müller and Fishel, 2002].

1.4.5.1 Homologous recombination (HR)

Homologous recombination (HR) is responsible for repairing double-strand breaks (DSBs), which occur if DNA is damaged in the S or G2 phase of the cell cycle, since there are two copies of each chromosome in this stage of the cycle. Of all repair pathways available for the repair of DSBs, HR is the least error-prone as it is template-based. DSBs are commonly caused by chemotherapy in cancer treatment.

Both BRCA1 and BRCA2, tumour suppressor genes linked to the development of inherited breast and ovarian cancer, have been found to be vital components of the homologous recombination machinery [Boulton, 2006, 2010]. 80% of people with germline mutations in BRCA1 or BRCA2 will develop breast cancer by the age of 70 years [Boulton, 2006]. BRCA2 is responsible for recruiting and loading the RAD51 repair protein

onto DSBs, whilst BRCA1 signals the presence of DSBs to the cell cycle, although its exact function in suppressing cancer development is unclear [Boulton, 2010]. It has been shown in mice knock-out models, that the deletion of BRCA1 does not allow the DSB to be processed and fixed by HR, allowing 53BP1 to inhibit DSB processing, contributing to tumourigenesis [Bouwman et al., 2010, Bunting et al., 2010]. However, this phenotype is rescued by also knocking out the 53BP1 protein in BRCA1-deficient cells, which relieves the inhibition of DSB processing, so homologous recombination is restored and the risk of developing cancer is suppressed once more. This finding is very promising in terms of personalised medicine, as it has been proposed by Bunting et al. [2010] that 53BP1 inhibitors could be used to treat carriers of BRCA1 mutations, to suppress their susceptibility to tumour formation.

1.4.5.2 Classical non-homologous end joining (C-NHEJ)

Classical non-homologous end joining (C-NHEJ) is the only double strand break repair pathway that can join DNA ends with no homology at the repair site [as reviewed in Bunting and Nussenzweig, 2013].

C-NHEJ involves the binding of the heterodimer Ku70Ku80 to the DNA break, followed by the recruitment of DNA-dependent protein kinase catalytic subunit (DNA-PKcs) and several other factors that mediate blunt-end ligation of the break by DNA ligase 4 (LIG4) [as reviewed in Bunting and Nussenzweig, 2013].

NHEJ has been implicated as a key source of genomic rearrangements in cancer, especially complex chromosome translocations that are associated with chromothrypsis [as reviewed in Bunting and Nussenzweig, 2013].

1.4.5.3 Alternative end-joining (A-EJ)/ Microhomology-mediated end-joining (MMEJ)

Alternative end-joining (A-EJ) uses microhomology to fix DSBs, and has been shown to be the compensatory mechanism in HR-deficient cancers when homologous recombination is defective. However, like C-NHEJ, A-EJ is far more error-prone than HR, and so causes a higher mutation rate [Ceccaldi et al., 2015]. POLQ has been implicated in this pathway, shown to have increased expression levels in HR-deficient cancers [Ceccaldi et al., 2015].

In cancer cells, A-EJ pathways join DSBs in aberrant ways which promote cancer growth. This pathway, like C-NHEJ, also gives rise to an increased rate of translocations [as reviewed in Bunting and Nussenzweig, 2013].

1.4.5.4 Nucleotide-excision repair (NER)

Nucleotide excision repair (NER) is another repair process that is disrupted in cancer [Leibeling et al., 2006]. NER removes primarily bulky, helix-distorting adducts. These substrates include *cis-syn*-cyclobutane dimers (CPDs) and pyrimidine (6-4) pyrimidone photoproducts, both of which are formed between adjacent pyrimidines and are induced by UV light [de Boer and Hoeijmakers, 2000]. Other NER substrates include bulky chemical adducts such as large polycyclic aromatic hydrocarbons induced by compounds in cigarette smoke; and distorting interstrand crosslinks induced by chemotherapeutic agents such as cisplatin [de Boer and Hoeijmakers, 2000].

Defects in NER are associated with the hereditary human disease Xeroderma pigmentosa [Cleaver, 2005], which is known to predispose to the development of skin cancer.

1.4.5.5 Base-excision repair (BER)

It is estimated that ~20,000 potentially mutagenic lesions arise per diploid mammalian cell per day as a result of intrinsic sources such as chemical instability of DNA under

physiological conditions and exposure of DNA to active oxygen. The majority of these mutations are repaired by base-excision repair (BER), which must occur efficiently prior to DNA replication in order to minimise spontaneous mutation rate [Preston et al., 2010]. BER is considered as the main pathway involved in removal of minor base damage induced by alkylating and oxidising agents, which are generally not helix distorting [de Boer and Hoeijmakers, 2000].

1.4.5.6 Mismatch repair (MMR)

Mismatch repair (MMR) is also a type of excision repair. MMR is a process by which base-base mismatches and insertion-deletion loops which arise as a consequence of DNA polymerase slippage during DNA replication are eliminated. However, in cancer this process is sometimes defective. Mutation rates in tumour cells with MMR deficiency are 100 to 1000-fold as compared with normal cells. For example, this is known to occur in hereditary non-polyposis colon cancer in MSH2 and MLH1 [Müller and Fishel, 2002, Peltomäki, 2001]. This mechanism is involved in the replication process, repairing base pair mismatches that have escaped polymerase proofreading during DNA synthesis. However, MMR machinery is also involved in apoptosis [Preston et al., 2010].

1.4.6 Structural genome instability

Genomic instability is characteristic of most human cancers [as reviewed in Negrini et al., 2010]. There are various forms of genomic instability: chromosome instability (CIN), microsatellite instability (MIN or MSI) and forms of genetic instability characterised by increased frequencies of base-pair mutations [as reviewed in Negrini et al., 2010]. It is well established that chromosome instability (CIN) and microsatellite instability (MIN or MSI) predispose to cancer [Preston et al., 2010], and in hereditary cancers is known to result from mutations in DNA repair genes [as reviewed in Negrini et al., 2010]. However in sporadic cancers, the molecular basis of genomic stability is less clear [as reviewed in Negrini et al., 2010].

1.4.6.1 Chromosome instability (CIN)

CIN is the major form of genetic instability in human cancers. This refers to the high rate by which chromosome structure and number changes over time in cancer cells compared to normal cells [as reviewed in Negrini et al., 2010].

In hereditary cancers, the presence of CIN has been linked to mutations in DNA repair genes. For example, germline mutations in genes associated with the repair of DSBs or DNA interstrand cross links such as BRCA1, BRCA2, PALB2, BRIP1, RAD50, NBS1, WRN, BLM and RECQL4 predispose to the development of various cancers including breast, ovarian, leukemia and lymphoma [as reviewed in Negrini et al., 2010].

CIN characterises almost all sporadic human cancers, however rather than mutations in DNA repair genes or other caretaker genes, it is possible that oncogene-induced DNA replication stress is responsible for this genomic instability in human cancers [as reviewed in Negrini et al., 2010]. According to this model, activated oncogenes induce genomic instability through the mechanism of DNA replication stress. This is supported by the findings that genomic instability preferentially affects common fragile sites which are known to be particularly sensitive to DNA replication stress [as reviewed in Negrini et al., 2010].

1.4.6.2 Microsatellite instability (MIN/ MSI)

Microsatellite instability (MIN) is a form of structural genome instability that is characterised by the expansion or contraction of the number of oligonucleotide repeats present in microsatellite sequences [as reviewed in Negrini et al., 2010]. These insertions and deletions determined by MSI are often >8bp.

MIN is a hallmark of MMR loss, particularly relevant for colorectal cancer because many of the genes involved in colorectal cancers have repetitive DNA in their coding regions (e.g. APC, KRAS, BRAF) [Preston et al., 2010]. One of the best documented

occurrences of MIN is in hereditary non-polyposis colon cancer, in which mutations in DNA MMR genes such as MSH2 lead to MIN [Fishel et al., 1994].

1.4.7 Genome editing

Genome editing is an approach used to generate desired genetic modifications, and is enabled by the induction of a DSB in a specific genomic target sequence. The modifications are then created during endogenous subsequent DNA break repair [Urnov et al., 2010]. This method exploits how DNA is damaged in cancer cells in order to treat cancer.

This process is carried out using a zinc finger nuclease (ZFN), a sequence specific endonuclease that can cleave the chosen target. ZFN-mediated gene disruption has been applied to cancer therapy, specifically in the treatment of glioblastoma. In this case the glucocorticoid receptor gene is disrupted.

The type II bacterial CRISPR(clustered, regularly interspaced, short pallindromic repeats)-Cas9(CRISPR-associated protein) system can be engineered to introduce RNA-directed DSBs using a mutated form of Cas9. CRISPR-Cas9-based genome editing enables the rapid genetic manipulation of any genomic locus without the need for gene targeting by homologous recombination, providing a simple strategy to develop conditional, 'deletion' models *in vivo* in mice in less than six months. This is a flexible, fast and low-cost platform to study gene function [Dow et al., 2015].

1.5 Detecting mutations

A central goal of cancer genetics is to identify cancer genes carrying driver mutations, and distinguish those drivers from the more abundant but inconsequential passenger mutations that are present in tumours. Many studies have attempted to do this; such work has provided insights into the genetics of cancer initiation and progression.

Traditionally, linkage and association studies were used to identify heritable cancer predisposing mutations (e.g. the germline mutation BRCA2 which predisposes to breast and ovarian cancer). For the identification of somatic changes a candidate screening approach was typically adopted, in which candidate genes were PCR amplified and sequenced in tumours and their matched somatic tissue using capillary Sanger sequencing [Bignell et al., 2006]. These candidates came from identified germline genes in linkage and association studies as well as being based on functional candidacy (e.g. a gene that is in the same pathway as a known germline mutated cancer gene, such as BRCA2, is a candidate for these somatic mutation analysis studies). Using a candidate approach, the first driver mutation in cancer was identified around 1980 in codon 12 of the oncogene HRAS; a G>T substitution causing a glycine to valine change [Reddy et al., 1982, Tabin et al., 1982]. However, this approach is limited to screening what is known, and does not have the ability to discover novel cancer genes.

The field of cancer genomics is, however, now in the process of being transformed by advances in DNA sequencing technology, and with the recent advent of more economical methods such as next-generation sequencing it is possible to sequence a whole cancer genome to detect somatic cancer-causing mutations. This is much faster and cheaper than previously used traditional Sanger sequencing techniques. While these next-generation sequencing prices are falling it is still expensive to sequence a whole cancer genome, however we can now use established techniques such as targeted exome capture to sequence just the exomes of cancer genomes.

1.5.1 DNA sequencing technologies

DNA sequencing is the process of determining the precise order of the four bases in the genome: adenine (A), cytosine (C), guanine (G) and thymine (T).

1.5.1.1 Sanger sequencing

Since the early 1990s, DNA has been sequenced using capillary-based, semi-automated implementations of the Sanger biochemistry (Figure 1.7).

DNA to be sequenced is first prepared in one of two ways: for shotgun *de novo* sequencing, randomly fragmented DNA is cloned into a high-copy-number plasmid which is then used to transform *Escherichia coli* (shown in Figure 1.7); or for targeted resequencing (in cases where a reference sequence is available; the approach adopted after the completion of the human genome sequencing project), PCR amplification is carried out with primers that flank the target. This results in amplified templates of the DNA sequence to be sequenced, either in the form of many clonal copies of a single plasmid insert present within a spatially isolated bacterial colony that can be picked, or as many PCR amplicons present within a single reaction volume [as reviewed in Shendure and Ji, 2008].

Sequencing takes place as a 'cycle sequencing' reaction, in which cycles of template denaturation, primer annealing and primer extension are performed. The primer is complementary to the flanking sequence of the region of interest. For each round of the primer extension, deoxynucelotides (dNTPs) are added to the reaction and polymerase is used to synthesise a new strand complementary to the template. The primer extension is randomly terminated by the incorporation of fluorescently-labelled dideoxynucleotides (ddNTPs), with each of the four different ddNTPs (ddATP, ddCTP, ddGTP and ddTTP) carrying a different fluorescent label. ddNTPs differ from dNTPs by the lack of a free 3' OH group on the five-carbon sugar, so if a ddNTP is added to a growing DNA strand then the chain is terminated as there is no free 3' OH group available for the addition of another dNTP. In the resulting mixture of end-labelled extension fragments, the label on the terminating ddNTP of any given fragment therefore identifies the nucleotide at the terminating position [as reviewed in Shendure and Ji, 2008].

The sequence is then determined using high-resolution electrophoretic separation of the single-stranded fragments in the mixture of end-labelled extension products, in a

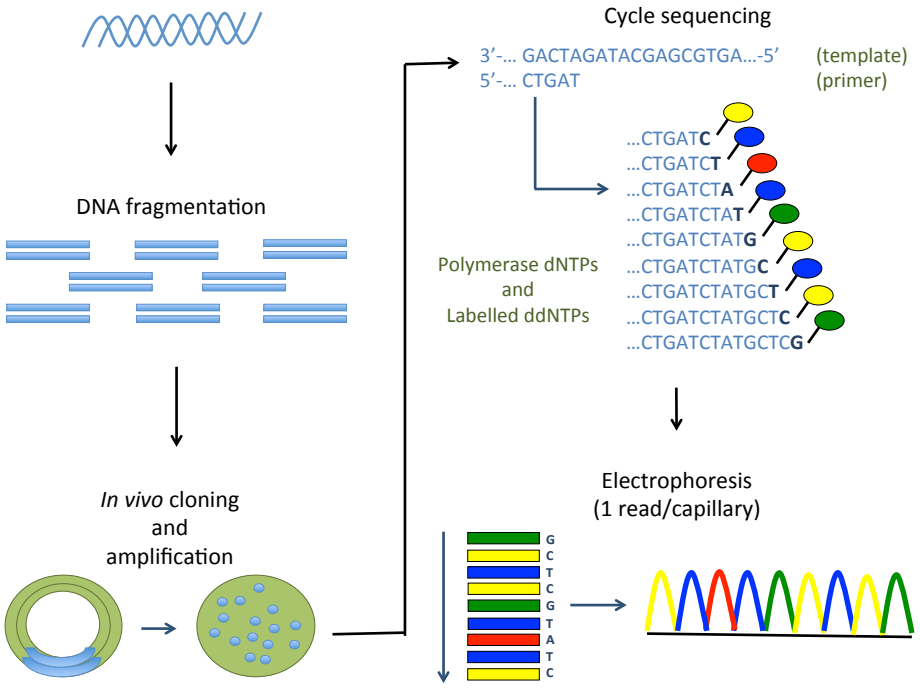


FIGURE 1.7: **Sanger capillary sequencing.** In high-throughput shotgun Sanger sequencing, DNA is fragmented, cloned to a plasmid vector and used to transform *E. coli*. Then for each cycle sequencing reaction, a single bacterial colony is picked and the plasmid DNA is isolated, before a ladder of ddNTP-terminated, dye-labelled products is generated (within a microliter-scale volume). These are subjected to high-resolution electrophoretic separation within one of 96 or 384 capillaries in one run of a sequencing instrument. The fluorescently labelled fragments of discrete size pass a detector, generating a sequence trace using the four-channel emission spectrum. *Adapted from Shendure and Ji [2008].*

capillary-based polymer gel, to separate the fragments by size. This can be carried out simultaneously in 96 or 384 independent capillaries, providing a limited level of parallelisation. Laser excitation of fluorescent labels as fragments of discreet lengths exit the capillary, together with four-colour detection of emission spectra, provides the readout as a Sanger sequencing 'trace' [as reviewed in Shendure and Ji, 2008].

1.5.1.2 Next-generation sequencing

Next-generation sequencing (NGS) instruments provide higher throughput than traditional Sanger sequencing at an unprecedented speed by sequencing millions of short

DNA fragments in parallel (Figure 1.8). The various platforms used to perform next-generation sequencing include: 454, Illumina (Solexa), SOLiD and Polonator [as reviewed in Shendure and Ji, 2008]. The most widely used platform is Illumina, launched in 2006 [Pabinger et al., 2014]. Illumina sequences DNA by measuring and analysing signals which are emitted during the creation of the second DNA strand. In order to produce detectable signals, template DNA is fragmented into small pieces, amplified and immobilised on a glass slide before sequencing. Illumina applies a sequencing-by-synthesis approach where only 1nt per sequencing cycle is incorporated using reversible dye terminators [Pabinger et al., 2014]. This particular platform has the benefit of being able to avoid problems when sequencing homopolymers (consecutive instances of the same base, eg. AAA or GGG) longer than 8bp encountered by other platforms such as Roche 454, which makes Illumina more suitable for identifying INDELs. However this comes at the cost of only being able to sequence shorter fragments in Illumina [Pabinger et al., 2014].

As in Sanger sequencing, the first step in next-generation (cyclic-array) sequencing is the preparation of a library containing DNA templates for sequencing. This is accomplished by the random fragmentation of DNA, followed by *in vitro* ligation of common adapter sequences [as reviewed in Shendure and Ji, 2008].

The generation of clonally clustered amplicons to serve as sequencing features is then achieved through various approaches depending on the platform. Illumina relies on bridge-PCR (aka 'cluster PCR') to clonally amplify sequencing features, whereas 454, SOLiD and Polonator all use emulsion PCR. In the bridge PCR approach adopted in Illumina platform (Figure 1.9), both forward and reverse PCR primers are tethered to a solid substrate by a flexible linker, so that all amplicons arising from any single template molecule during the amplification remain immobilised and clustered to a single physical location on the array. Extension is carried out using a *Bst* polymerase, and the double stranded sequence is denatured using formamide. The resulting clusters of amplified template consist of ∼1,000 clonal amplicons, and several million clusters can be amplified to specific locations within each of the eight lanes on a single flow-cell.

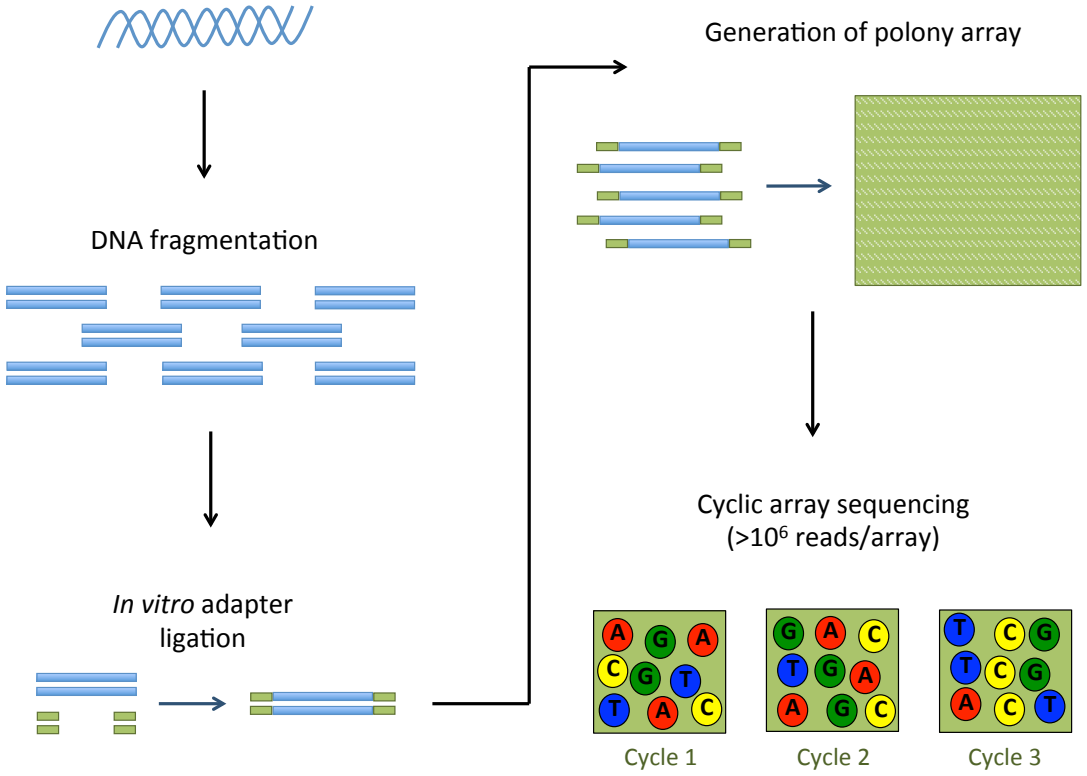


FIGURE 1.8: **Next-generation sequencing.** In cyclic-array shotgun sequencing, common adaptors are ligated to fragmented DNA. An array of millions of spatially immobilised PCR colonies ('polonies') are then generated using various different methods (e.g. bridge PCR used by the Illumina platform), with each polony consisting of many copies of a single shotgun library fragment. All array features are then sequenced in parallel, using primer hybridisation, enzymatic extension reactions and imaging-based detection of fluorescent labels incorporated with each extension. A contiguous sequencing read for each array feature is built up using successive iterations of enzymatic interrogation and imaging. *Adapted from Shendure and Ji [2008].*

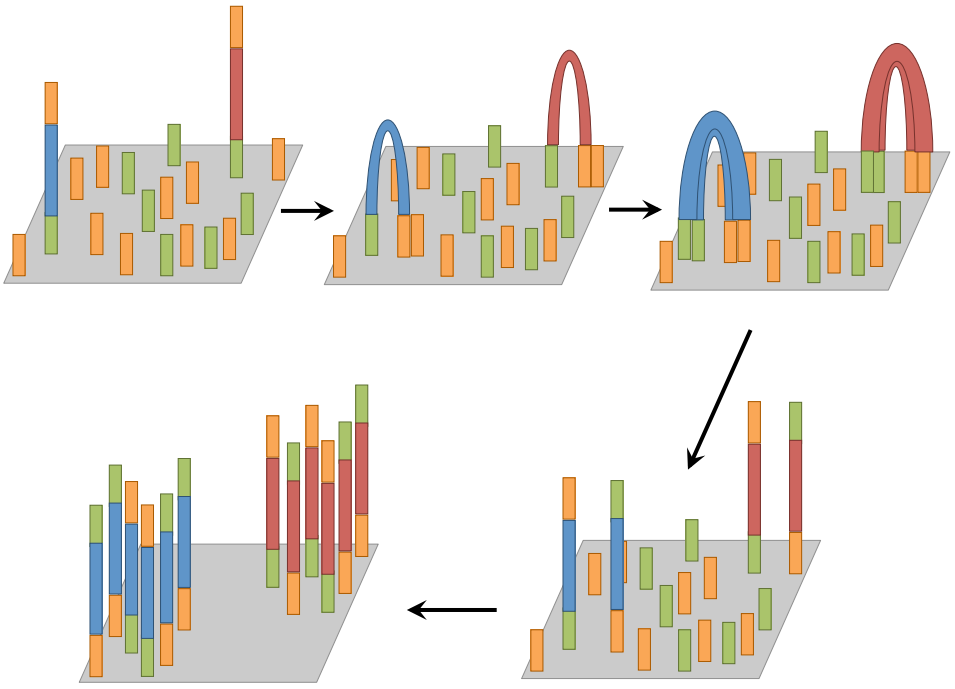


FIGURE 1.9: **Illumina bridge PCR.** Using the Illumina Genome Analyser ('the Solexa') as a platform, amplified sequencing features are generated by bridge PCR. The resulting clusters consist of ~1000 clonal amplicons. An *in vitro*-constructed adaptor-flanked shotgun library is PCR amplified, with both primers attached at their 5' ends by a flexible linker densely coating the surface of a solid substrate (flow-cell). As a consequence of this set-up, amplification products originating from a single member of the template library will remain locally tethered near the point of origin. *Adapted from Shendure and Ji [2008].*

Therefore, eight independent libraries can be sequenced in parallel in the subsequent step, during the same instrument run [as reviewed in Shendure and Ji, 2008].

For the sequencing step, 454 and Illumina both use polymerase for their sequencing-by-synthesis with 454 using pyrosequencing and Illumina using reversible terminators (Figure 1.10). Both SOLiD and Polonator use ligase to synthesise with SOLid using octamers with two-base encoding and Polonator using nonamers to do this [as reviewed in Shendure and Ji, 2008]. In Illumina sequencing, after cluster generation, the amplicons are single stranded by linearisation and a sequencing primer is hybridised to a universal sequence immediately flanking the region of interest. Each cycle of sequence interrogation then consists of single-base extension with a modified DNA polymerase and a

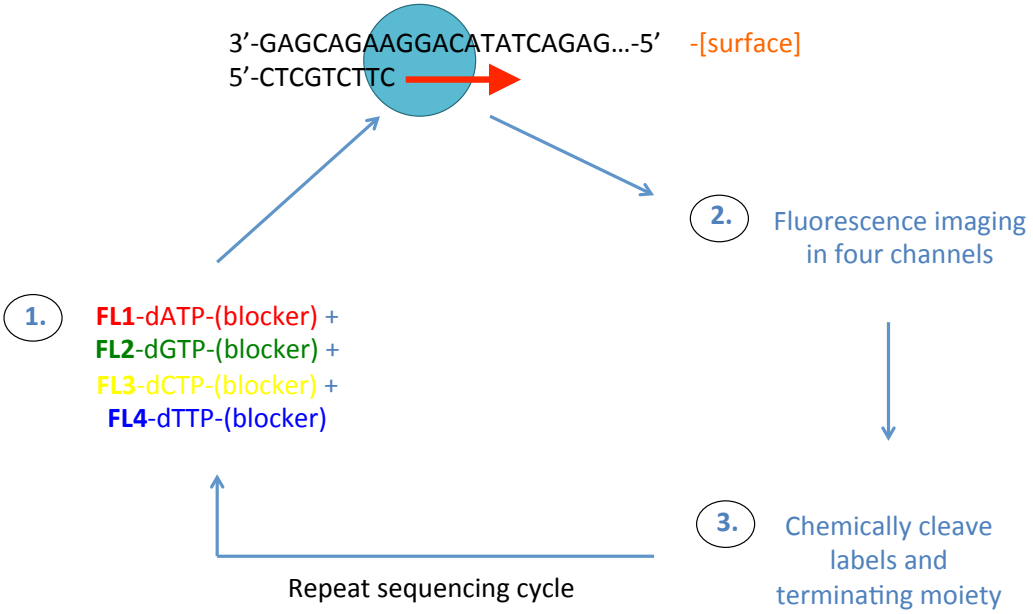


FIGURE 1.10: **Illumina sequencing.** With the Solexa technology, each sequencing cycle includes the simultaneous addition of a mixture of four modified deoxynucleotide species each bearing one of four fluorescent labels and a reversibly terminating moiety at the 3' hydroxyl position. A modified DNA polymerase drives simultaneous extension of primed sequencing features. This is followed by imaging in four channels and then cleavage of both the fluorescent labels and the terminating moiety, before the next cycle. *Adapted from Shendure and Ji [2008].*

mixture of four fluorescently labelled nucleotides, with each chemically cleavable label corresponding to the identity of the nucleotide. These modified nucleotides are also 'reversible terminators' with a chemically cleavable moiety at the 3' hydroxyl position, which allows only one base to be incorporated in each cycle. After single-base extension and acquisition of images in four channels, chemical cleavage of both groups sets up for the next cycle, hence why the nucleotides are known as reversible terminators [as reviewed in Shendure and Ji, 2008].

1.5.1.3 Targeted exome capture and sequencing

Exome sequencing (exome-seq) is performed using exome capture followed by next-generation sequencing, to capture reads covering the exonic regions of the human

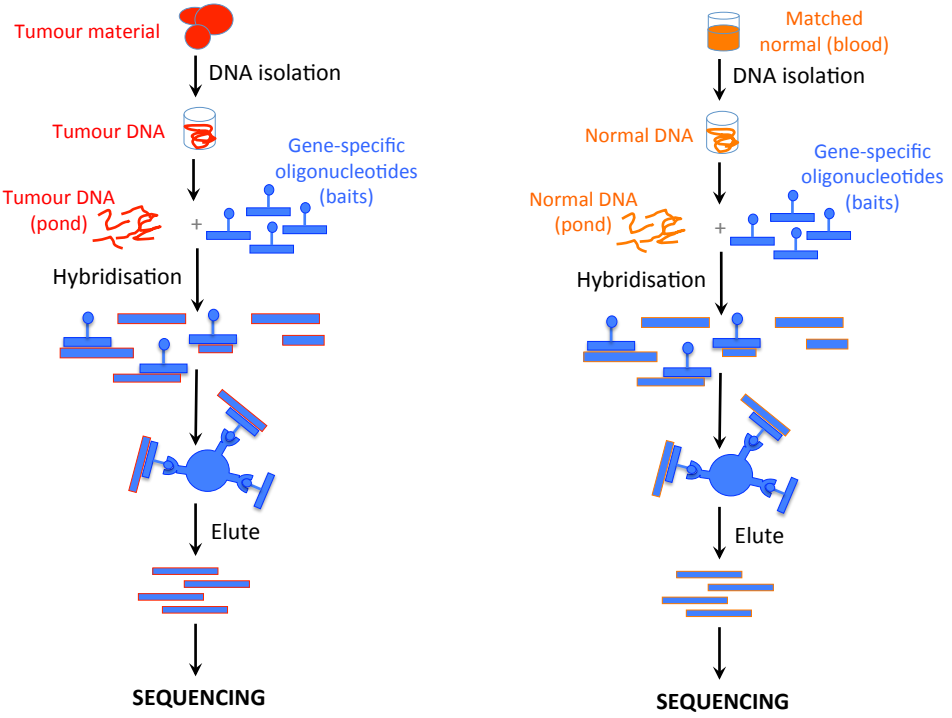


FIGURE 1.11: **Targeted exome capture.** This approach can be applied to any region of interest, but in the case of targeted exome capture the gene-specific oligonucleotides capture just the exons in genes in the whole exome. Exons are captured from both tumour DNA (left panel) and normal DNA (right panel). DNA from the starting material (pond) is sheared and hybridised to the oligonucleotides that are specific to exons only. The baits have tags that allow them to be isolated, by immobilisation on beads for example as shown here. Captured DNA in then eluted and prepared into sequencing libraries for sequencing. *Adapted from Meyerson et al. [2010].*

genome only. This is done using hybridisation to target the exome sequences (Figure 1.11). The exome encompasses the protein-coding regions of the genome, which consists of the exons of genes only. The remainder of the genome, including the introns of genes, is not translated into proteins and therefore is less likely to contain functionally important driver mutations. Exome sequencing is also currently much more economical than whole-genome sequencing, in terms of cost and time, since the exome only constitutes $\sim$ 1% of the genome [Chilamakuri et al., 2014, Meynert et al., 2013, Rabbani et al., 2014]. For these reasons, cancer studies have focused mainly on detecting mutations in whole-exomes rather then whole-genomes.

Target sizes of 35 to 50Mb are standard in exome capture, compared to the 3200Mb size

of the haploid genome. A typical targeted exome would encompass both protein-coding regions, other functional regions of interest such as promoters, enhancers and regulatory RNA genes, as well an some intronic regions. It is possible to see intronic and other non-coding variants in exome capture data for three reasons: probe design, alternative splicing and non-specific target capture [Meynert et al., 2013]. In targeted sequence capture, the probe is designed to have a minimum length of 120-200bp. This means that if an exon is shorter than the probe, some of the intronic sequences surrounding the exon will also be captured. Alternatively, if the exon is larger than the probe, then several probes will be needed to capture the exon, which also results in intronic variants being inadvertently captured. Often, depending on the kit being used, a proportion of the downstream splice junction is also captured (up to 10bp). Alternative splicing can also play a role, since genes with multiple transcript types will have a particular exon present in one transcript but not in another. Therefore variants in that exon can be considered exonic in some cases and intronic in others. Finally, it is possible that up to 50% of the sequence reads can be off-target, aligning to introns instead of exons.

A caveat of exome sequencing is that some exome-seq targets do not achieve sufficient mapped read depth for variant detection, due to technical difficulties or probe failures. This is compared to whole-genome sequencing which has a greater uniformity of sequence read coverage and reduced biases in the detection of non-reference alleles [Meynert et al., 2014].

1.5.1.4 Whole-genome sequencing

Whole-genome sequencing provides a complete view of the human genome, including the ability to detect mutations in distant enhancers and other regulatory elements such as promoters, and other non-coding regions such as introns and non-coding RNAs (including microRNAs), that are not captured in whole-exome sequencing [Pabinger et al., 2014].

The major potential of whole-genome sequencing for cancer is in the discovery of chromosomal rearrangements which cannot easily be identified using exome sequencing, including rearrangements of repetitive elements (e.g. active retrotransposons which have been suggested to be involved in cancer) [as reviewed in Meyerson et al., 2010].

1.5.1.5 The challenge of tumour heterogeneity

The cellular heterogeneity of tumours poses a major challenge for the reliable detection of genetic changes. The primary (solid) tumour is invariably infiltrated by a mixed population of non-cancerous cells, for example macrophages, endothelial cells and normal cells from the tissue of origin [as reviewed in Meyerson et al., 2010]. Cancer sequencing projects often attempt to minimise such issues by requiring tumour samples sent for sequencing to contain >80% abnormal cells as scored by a pathologist [Cancer Genome Atlas Research Network, 2008].

The second challenge of heterogeneity is that there are often sub-populations within the wider population of cancer cells in a tumour [Cleary et al., 2014]. Measured sub-populations can harbour distinct mutations, varying between regions of the tumour, between metastases and over time [Burrell et al., 2013]. Assuming (as is typically assumed) that the cancerous cells of a tumour are monoclonal in origin, then the mutations leading to the development of the cancer and the earliest mutations post-transformation are likely to be common to all cancerous cells in the tumour (excepting the possibility of subsequent deletion). Most recent mutations however are likely to be present at a lower frequency in the analysed tumour sample and are challenging to discriminate from assaying or sequencing errors. Despite these challenges arising from heterogeneity, the alternative is to use tumour derived cell lines which is potentially more problematic, as the cells that grow from a tumour in culture may be rather unrepresentative of the main cancer population within the tumour.

The practical consequence of tumour heterogeneity is that standard variant detection and genotype calling as applied in germline genetic analysis is not immediately applicable to primary tumour analysis and modified assumptions and pipelines are required.

1.5.1.6 Depth of coverage and physical coverage

Depth of coverage is measured by the amount of over-sampling, representing the number of sequenced reads that cover the site [as reviewed in Meyerson et al., 2010]. Sequence coverage affects the ability to detect point mutations. Most whole-exome sequencing studies aim for an average of 100 to 150-fold coverage, and whole-genome sequencing studies aim for about 30 to 60-fold coverage, in order to reach deep coverage. This is necessary owing to the heterogeneous nature of most tumours, which reduces the sensitivity to detect mutations, and genomic regions of high GC content which typically result in low or absent coverage [as reviewed in Watson et al., 2013].

There is considerable heterogeneity of target coverage encountered with exome sequencing technology, which can systematically affect the sensitivity to detect genuine mutations. This can be accounted for by applying an empirical model relating the read depth at a polymorphic site to the probability of calling the correct genotype at that site which quantifies the amount of variation missed at a given coverage threshold, as has been done in Meynert et al. [2013].

Physical coverage measures the number of fragments that span the site, and is based on using paired-end or mate-pair reads, making it possible to sequence reads that are of known chromosomal distance. This is an important consideration when detecting rearrangements, since paired reads provide additional information to enhance mapping accuracy of structural variants [Pabinger et al., 2014]. The expected distance between the paired reads is used to uniquely place the reads on the reference genome, with unexpected placement of read pairs can be used to detect a structural variation [as reviewed in Meyerson et al., 2010].

1.5.2 NGS variant analysis pipelines

The basic workflow of whole-exome and whole-genome sequencing projects is displayed in Figure 1.12. After sequencing, a bioinformatic analysis pipeline must be employed to successfully handle and analyse the huge amount of raw sequencing data output from next-generation sequencing.

1.5.2.1 Quality assessment

The standard file format for NGS raw data is FASTQ, which is a text-based representation of biological sequences beginning with the sequence name followed by lines of single-letter coded nucleotides or amino acids, and Phred-scaled base quality scores to facilitate the assessment of sequence quality [as reviewed in Bao et al., 2014].

The first analysis step after sequencing is to assess the quality of the raw sequencing reads, preprocessing raw data before alignment. This involves removing, trimming and correcting reads that do not meet defined standards, dependent on base quality scores and sequence properties (e.g. primer contaminations, N content and GC bias) [Pabinger et al., 2014]. Sequence artefacts that may compromise the quality of the raw sequencing data include: base calling errors, INDELs, poor quality reads and adaptor contamination [Pabinger et al., 2014]. Standard preprocessing procedure therefore includes: 3' end adaptor removal, trimming of low quality bases at the ends of reads, and read filtering by removing undesired sequences such as contamination from primers and adaptors [as reviewed in Bao et al., 2014].

1.5.2.2 Alignment

Reads are aligned to a reference genome after they have been preprocessed to meet a certain quality standard. Some examples of alignment programs are BWA, Stampy, Bowtie/Bowtie2 and Novoalign. First-generation short-read aligners were optimised for ungapped alignments, whereas more recent programs are better at dealing with longer

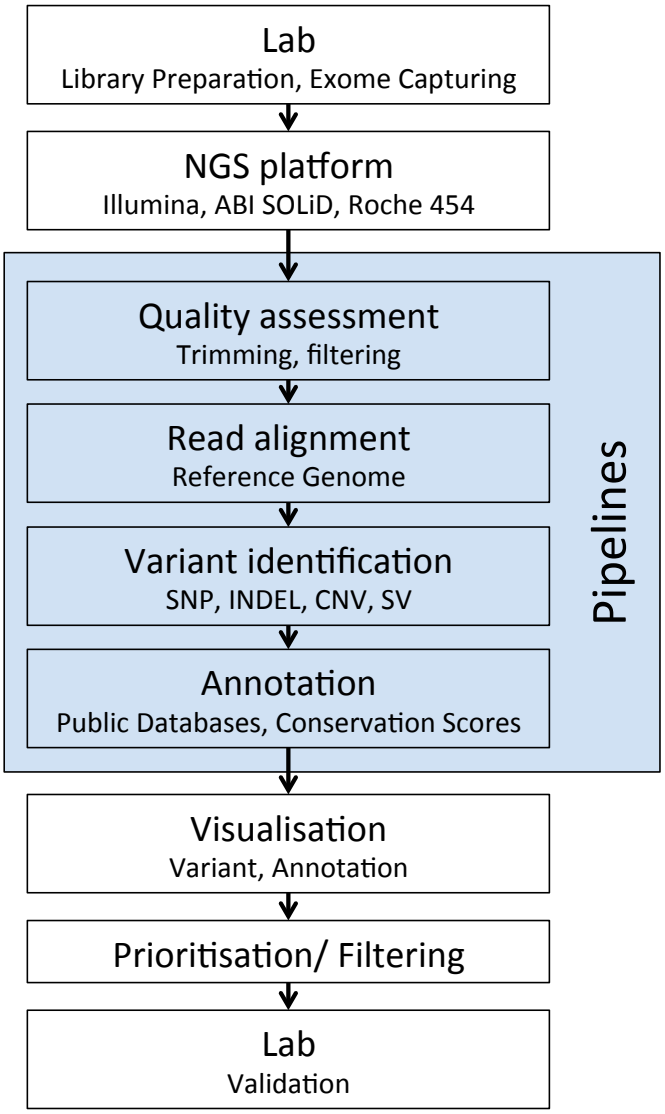


FIGURE 1.12: **NGS variant analysis workflow.** Whole-exome-seq or whole-genome-seq samples are prepared in a library before being sequenced on a certain platform. A pipeline follows consisting of the following steps: (i) quality assessment of raw data, (ii) read alignment to a reference genome, (iii) germline and somatic variant identification and (iv) annotation of detected variants to infer the biological relevance. Pipeline results can be displayed using specific data visualisation tools such as circos plots and genome browsers. Mutations can then be further filtered and prioritised to find potential driver mutations, followed by validation in the lab. *Adapted from [Pabinger et al., 2014].*

read lengths and gaps allowing imperfect matches [as reviewed in Bao et al., 2014]. Current long-read alignment algorithms are classified as either using hash table indexing or using compressed tree indexing based on Burrows-Wheeler transform [Pabinger et al., 2014].

The use of paired-end reads in sequencing steps help in overcoming the problem of ambiguity when mapping short reads to a reference genome [Pabinger et al., 2014], since it is possible that a short-read sequence can be matched to multiple locations in the reference genome if not using paired-end reads.

It is common practise to remove PCR duplicates after alignment in a further post-alignment quality assessment step. PCR duplicates are a result of the PCR steps in library preparations, which in turn can sometimes cause multiple reads that originate from only one template being sequenced and interfere with variant calling statistics [Pabinger et al., 2014].

1.5.2.3 Variant identification

Tools for genome-wide variant detection can be classified as either: (i) germline callers, (ii) somatic callers, (iii) CNV identification and (iv) identification of other SVs including translocations, inversions and large INDELs. However, cancer studies focus on detecting somatic mutations by comparing sequences of tumour/normal pairs from one patient. CNVs are the only large structural modification that can be detected in both whole-exome and whole-genome sequencing studies [Pabinger et al., 2014].

Multiple sample variant calling is usually recommended, in which all reads across multiple samples from one genomic location are taken into account. This approach reduces the possibility of calling randomly presented sequencing errors and increases the probability of calling alleles of low frequency or low coverage in a single sample. This increases the accuracy and sensitivity of multi-sample variant calling, compared to that of single sample variant calling. However, multi-sample variant calling is not always feasible with very large sample sizes [as reviewed in Bao et al., 2014].

1.5.2.4 Variant annotation

The functional impact of the identified mutations must be predicted in order to help distinguish the potential driver mutations from the passenger mutations, by enabling the downstream filtering and prioritisation of potential disease-causing variants for further analysis in the lab. Most annotation tools focus on SNVs as they are more easily identified and analysed than other forms of variation, although INDELs are also covered by some tools. Annotation of structural variants is limited to CNVs. The most commonly used public variant database used for this analysis step is dbSNP [Pabinger et al., 2014].

Examples of software designed to annotate mutations with their predicted effects on genes and protein function include Ensembl variant effects predictor [McLaren et al., 2010] and snpEff [Cingolani et al., 2012]. Loss-of-function and gain-of-function consequences of mutations can also be predicted using PolyPhen and SIFT [Flanagan et al., 2010].

1.5.3 Cancer whole-exome and whole-genome next-generation sequencing projects

Ng et al. [2009] were the first to use targeted exome capture sequencing. By comparing mutations in the exomes of control individuals and four patients with Freeman-Sheldon syndrome (FSS), they demonstrated the re-discovery of mutations responsible for a Mendelian trait. Adapting that approach to find the somatic mutations that underlie the development and progression of cancer, Timmermann et al. [2010] captured and sequenced tumour and matching normal colon tissue in the first study to work on whole exome next-generation sequencing of primary colon cancers.

Pleasance et al. [2010a] were the first to apply whole-genome next-generation sequencing to cancer, identifying somatic mutations in a whole malignant melanoma genome, in the first comprehensive catalogue of somatic mutations from an individual cancer. This

study also provides first insights into genomic alterations induced by ultraviolet light exposure.

Major large-scale cancer sequencing projects are now underway, such as the Cancer Genome Project (CGP) [Futreal et al., 2004] launched in the United Kingdom in 2000, The Cancer Genome Atlas (TCGA) [Cancer Genome Atlas Research Network, 2008] set up in the United States in 2006 and the International Cancer Genome Consortium (ICGC) [Zhang et al., 2011] created in 2007 to coordinate the generation of comprehensive catalogues of genome alterations from 52 different cancer types [as reviewed in Watson et al., 2013]. These initiatives aim to systematically decipher and catalogue the spectrum of genetic variants in different cancer types using next-generation sequencing.

Glioblastoma (GBM) was the first cancer type to undergo comprehensive genomic characterisation by the TCGA Research Network [Cancer Genome Atlas Research Network, 2008], in which a targeted approach was adopted using Sanger-based capillary sequencing of 601 selected genes in 91 tumour-normal exomes to identify somatic mutations. This analysis revealed frequent mutations in the phosphatidylinositide 3-kinase (PI3K) regulatory subunit PIK3R1.

The aim of the ICGC was to comprehensively characterise somatically acquired genetic events in at least fifty classes of cancer, including those with the highest global incidence and mortality, requiring the high-coverage sequencing of 20,000 cancer genomes or more. The catalogues of somatic mutation from these cancers are combined with expression and epigenetic profiles as well as clinical features from the same patients. This project builds on the success of previous collaborative initiatives such as the Human Genome Project and the HapMap consortium [as reviewed in Stratton et al., 2009].

This field has primarily focused on finding biomarkers (genes and pathways), for example APC, which can be used to guide therapy. These are mostly from exome (disease centric) studies, however many more whole genome sequencing studies are now starting to come out. For example Waddell et al. [2015] have used whole-genome sequencing and copy number variation (CNV) analysis of 100 pancreatic ductal adenocarcinomas

(PDACs) to uncover mutational mechanisms and copy number changes. A high prevalence of chromosomal rearrangements was discovered, with genomic instability found to co-segregate with the inactivation of DNA maintenance genes (BRCA1, BRCA2 or PALB2) and a mutational signature of DNA damage repair deficiency. These particular mutations would be difficult to capture with whole-exome sequencing, however they can be with whole-genome sequencing.

1.5.3.1 TCGA Pan-Cancer analysis project

The Pan-Cancer analysis project was launched by The Cancer Genome Atlas (TCGA) in 2012 to identify and analyse aberrations in the tumour genome and phenotype that define cancer lineages [Weinstein et al., 2013]. The main aim was to define commonalities, differences and emergent themes across different tumour types by assembling and comparing coherent and consistent TCGA datasets across twelve different tumour types (GBM, OV, BRCA, LUSC, LUAD, COAD, READ, KIRC, UCEC, BLCA, HNSC and LAML) and six different genomic, epigenomic, tanscriptional and proteomic platforms (mutation, copy number, gene expression, DNA methylation, microRNA and reverse phase protein array) combined with clinical data, profiled over 5,074 samples. Tumour types were selected based on data maturity, adequate sample size and publication of primary analyses. Since the launch of the Pan-Cancer initiative, the integrated dataset has undergone quality control, statistical analysis and interpretation by a consortium of researchers.

The purpose of the project is to increase statistical power through increased sample size to detect new patterns of functional genomic determinants of disease by using a collated dataset, increasing the power to also detect rare genomic drivers in heterogeneous tumours, as well as eliminating false-positives commonly seen in single tumour type analyses. Since cancers from disparate tumour types have been shown to share certain features, and cancers from the same organ have also been revealed to sometimes be quite distinct, another imperative objective of this project is to uncover both tissue-specific aspects of cancer as well as intrinsic molecular commonalities across many different

tumour types. This addresses the role of mutational profile, as well as tissue of origin, as an important determinant in the progression of cancer. The ultimate goal is that the results from this project will help to inform clinical decision making for individual cancers.

Recently published studies that have used the Pan-Cancer data in order to comprehensively identify candidate cancer genes over multiple cancer types include Lawrence et al. [2014] and Kandoth et al. [2013], in which detected driver mutations have been related to both cancer type and mutation spectra.

1.5.3.2 Distinguishing drivers from passengers

Many types of mutation have been uncovered using next-generation sequencing projects, for example small-scale re-arrangements [Stephens et al., 2009], large-scale genomic rearrangements [Stephens et al., 2011], loss of heterozygosity (LOH) [Zhao et al., 2010] and homozygous deletions [Bignell et al., 2010]. Segmental duplications and aneuploidy (abnormal number of chromosomes) have also been discovered in cancer using these methods. For example a tandem duplication was identified in a small-cell lung cancer genome in Pleasance et al. [2010b], shown to duplicate exons 3-8 of CDH7 in frame, and in Bignell et al. [2006] testicular germ-cell tumours were found to be uniformly aneuploid in protein kinase genes, with consistent chromosomal gains on 12p, 8q, 7, and X and losses on 13q, 18q, 11q, and 4q. The oncogene TRRAP was also detected through the next-generation sequencing of 14 melanoma exomes in Wei et al. [2011], by uncovering a recurrent cytosine to thymine mutation. Puente et al. [2011] also set out to identify new driver mutations in a genome-wide screen of chronic lymphocytic leukemia (CLL), identifying the oncogenes NOTCH1, MYD88 and XPO1, which provided a very interesting insight into the usefulness of combining whole-genome sequencing with clinical data in identifying driver mutations in a study that was the first of its kind.

Although these next-generation sequencing studies have proven to be very informative in identifying driver genes, they highlight limitations associated with these methods.

Driver mutations have several characteristics that separate them from passenger mutations; for example driver mutations tend to cluster in cancer genes whereas passengers are more randomly distributed [as reviewed in Stratton et al., 2009]. However, the overly simplistic approach that many studies have adopted of classing genes with an excess of non-synonymous mutations or frequently occurring (recurrent) mutations as driver genes, sometimes results in mistaking passengers for driver. These studies interpret an excess of observed non-synonymous mutations compared with that expected by chance as evidence for the presence of driver mutations, but do not include the examination of synonymous mutations. It is therefore assumed in these studies that the mutation rate across the genome is constant. However, mutation rate is known to vary across the genome with hotspots experiencing higher rates of mutation. So for example, in these studies a highly mutable locus could look like a cancer gene when in fact there is no guarantee that these mutations are drivers, as it is possible that the gene is being recurrently hit by passengers at a mutator hotspot. Therefore it is important that fluctuations in the regional mutation rate across the genome are accounted for [Yang et al., 2003].

Refined driver detection methods include using evolutionary conservation of a mutated residue, which can indicate the importance of the mutational event, helping to discriminate drivers from passengers. Algorithms such as Cancer-Specific High-Throughput Annotation of Somatic Mutations (CHASM) [Carter et al., 2009] are used to score mutational impact by amino acid conservation inferring probable functional mutations in cancer [as reviewed in Watson et al., 2013]. Somatic missense mutations are ranked using a many features classifier in order to identify driver mutations.

Vandin et al. [2012] also attempted to overcome the problems encountered by common approaches in cancer studies that predict driver mutations based solely on their frequency of occurrence. Since this approach is known to be confounded by the observation that driver mutations target multiple cellular signalling and regulatory pathways and thus each cancer patient may exhibit a different combination of mutations that are

sufficient to perturb these pathways, Vandin et al. [2012] have dealt with the mutational heterogeneity problem by applying an algorithm called Dendrix, to find driver pathways *de novo* from somatic mutation data.

1.6 Cancer as an evolutionary process

The development and adaptation of cancer can be viewed as an evolutionary process, and can therefore be defined in these terms, occurring through cell-level selection in much the same way that a species evolves through Darwinian organism-level natural selection [Talavera et al., 2010]. At least for the short term the formation of cancer is a case of survival of the fittest [Darwin and Wallace, 1858].

Thus driver mutations can also be explained in this context. Driver mutations have previously been described as mutations that are responsible for the development of cancer. This is because they confer growth, developmental and survival advantages on the cell in which they occur, and are therefore positively selected in the microenvironment of the tumour tissue through adaptive molecular evolution. Thus cells containing driver mutations will be best adapted for cancer development, with the capability to proliferate and survive more effectively than their neighbours [as reviewed in Stratton et al., 2009], and therefore increase in frequency and subsequently go on to dominate the cell population, leading to the formation of a malignant tumour. For example, a driver mutation that inactivates a tumour suppressor gene will confer a selective advantage to the cancer, so that this cell is now better adapted to the needs of the cancer by, for instance, being able to proliferate faster than other cells in the population. Conversely, it is the passenger mutations present in cancer that do not contribute to oncogenesis, as they do not confer a clonal growth advantage to the cancer and therefore have not been selected [as reviewed in Stratton et al., 2009]. Passenger mutations are present in the genomes of cancer cells because somatic mutations without functional consequence often arise during cell division. Thus, cells that acquire driver mutations will already contain passenger mutations and will carry them through further

rounds of mitosis, so that all cells of the final cancer contain passengers that were originally present in an ancestor of the cancer cell when it acquired one of its drivers [as reviewed in Stratton et al., 2009]. As well as driver and passenger mutations, some somatic mutations may occur in the cancer genome that impair cell survival. These mutations will be subject to negative selection, since they are detrimental to the cancer, and therefore will be absent from the cancer genome [as reviewed in Stratton, 2011].

Unlike selection at the species level, selection in cancer cell populations is an example of clonal selection, with cancer arising as an abnormal clone of cells that expand as a result of somatically acquired mutations (Figure 1.13) [as reviewed in Stratton, 2011, Stratton et al., 2009]. Clonal expansion is triggered by an initial driver mutation which confers a selective advantage to the cell over other cells in the population, and so this cell increases in frequency relative to cells without the driver mutation. The addition of a second driver mutation to a single cell within this tumour population confers extra growth advantages causing the cell to out-compete its neighbours once more. This process of clonal expansion is driven by selective pressures from the tumour's micro-environment or the external environment. For example, therapeutic intervention with chemotherapy can act as a strong selective pressure for the expansion of resistant sub-clones which manifest as recurrences, while at the same time destroying some cancer sub-clones. Other selective pressures of the tumour micro-environment include those that favour, for example: the outgrowth of clones possessing immune-evasive phenotypes [de Miranda et al., 2012]; cells that proliferate faster than their neighbours; and cells better adapted to acquire blood carrying oxygen and nutrient supplies. Cancer is constantly evolving in response to these changing selective pressures in the tumour micro-environment as well as in response to competition with itself (other cancer cells) [Talavera et al., 2010]. This is an example of the "Red Queen Hypothesis" which describes co-evolutionary relationships between different species as constant arms races, with each species rushing to evolve an advantage over the other, suggesting that faster evolution is favoured [Damore and Gore, 2011]. In terms of cancer, the different species are the different cells in the tumour population. So cancer cells therefore have to continually evolve at a fast rate to keep up with the ever-changing environment, and succeed in dominating

the tumour population. The population of neoplastic cells present in the final cancer derives from a single normal cell, as is shown in Figure 1.13, however the evolutionary history of the cell lineages is very complex. This is due to the multiple waves of clonal expansion that are thought to be required for the generation of the dominant sub-clone that manifests as the symptomatic cancer [as reviewed in Stratton, 2011]. Each new clonal expansion is initiated by an additional driver mutation, so there will be cases where driver mutations have caused sub-clones to branch off from the current sub-clone and then have failed to out-compete the subsequently dominant sub-clone. Therefore in the final cancer, some of these minor sub-clones will still exist, while others will have been extinguished. It is also possible that a minor sub-clone has branched off the current dominant sub-clone, which in time will go on to out-compete the major sub-clone and dominate the tumour. This will be due to a change in the micro-environment, for example after the treatment of chemotherapy in which a minor sub-clone containing a resistance mutation will preferentially expand to become the new major resistant sub-clone.

Cancer development and progression is in principle an attractive model for the study of evolution, due to the differences seen between cell-level selection in cancer and organism-level selection in species evolution. These benefits include: a short generation time in cancer lineages, meaning that analysis can be carried out on a much shorter time-scale; a lack of meiotic recombination, which can be a confounding factor in evolutionary analyses of eukaryotes and viruses; increased somatic mutation rate in cancer cells, the merit of which is that DNA sequence diversity on which selection can act is increased, providing strong signals of sustained positive selection [as reviewed in Stratton, 2011]; and the constantly challenging and changing environment including competition with other cells in the cancer population, again maximising the ability to detect sustained positive selection.

Cancers can therefore be defined as clonal proliferations, that arise owing to mutations in a subset of genes that confer selective growth advantage on cells [Greenman et al., 2007].

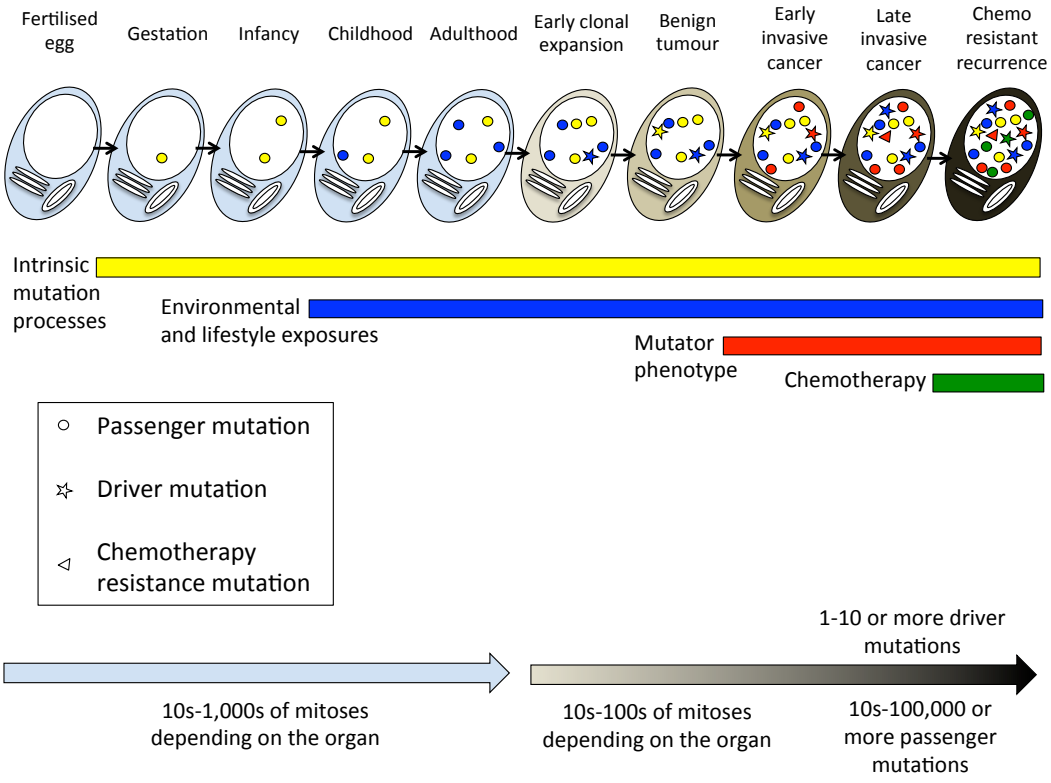


FIGURE 1.13: **The lineage of mitotic cell divisions in a cancer.** From the fertilised egg to a single cell within a cancer, somatic mutations may be acquired while the cell lineage is still phenotypically normal, represented by passenger mutations occurring before clonal expansion, caused by both intrinsic mutational processes during normal cell division and exogenous environmental mutagens. The rest of the somatic mutations in a cancer genome occur during the segment of the cell lineage in which the cancer cell is already showing phenotypic evidence of neoplastic change, and it is driver mutations that initiate this part of the lineage and continue to contribute to the clonal expansion of the cancer. Passenger mutations also accumulate during clonal expansion however they have no effect on the cancer cell. During clonal expansion processes such as DNA repair defects that cause a mutator phenotype contribute to the mutational burden. Chemotherapy can also contribute to the mutations seen in relapses after treatment, in which driver mutations confer resistance to cancer therapy and are likely to pre-date the initiation of treatment, existing previously as passenger mutations until the selective environment changed. *Adapted from Stratton et al. [2009].*

1.6.1 Detecting selection

The analogy between organism evolution and cancer progression has previously been recognised. One method that has been widely used in studies of selection during evolution is the "omega ratio", and this approach has also been applied to cancer data in order to distinguish driver mutations from passengers amongst all somatic mutations present in a cancer genome, and ultimately identify the cancer genes.

The omega ratio compares non-synonymous substitution rates (K_a) to synonymous substitution rates (K_s) to give a ratio of K_a/K_s in order to detect selection [Massingham and Goldman, 2005]. K_s acts as a proxy for neutral selection, since synonymous mutations do not alter the encoded protein and so are assumed to be selectively neutral. The synonymous rate is often used as a proxy for neutrality, which is the unbiased sampling of mutations (the effect of mutation and drift but not selection). Biological selection is hence expected to act only on non-synonymous mutations that alter the structure and function of proteins. The value obtained from this observed ratio indicates what type of selection is taking place: a value of <1 (a lower non-synonymous:synonymous ratio compared with that expected by chance alone) suggests evidence for negative selection overall; a value equal to 1 suggests no selection; and a selection pressure >1 (a higher ratio of non-synonymous:synonymous mutations compared with what is expected by chance) is indicative of positive selection, which is evidence for adaptive molecular evolution favouring advantageous driver mutations [Greenman et al., 2007]. The omega ratio is the comparison of two rates, so if the synonymous mutation rate is assumed to represent neutrality and the non-synonymous mutation rate represents neutrality plus some effect of selection, then the difference represents the net effect of selection, in this case positive selection.

This is a well established technique in the field of molecular evolution, used to detect selection acting on driver mutations in cancer. The benefit of using this approach over previously used methods of cancer gene detection is that it accounts for the varying regional mutation rate across the genome by using synonymous mutations as a control to estimate a rate rather than a count. In the past cancer genes (those containing

driver mutations) have been identified based on the presence of excess non-synonymous mutations or recurrently occurring mutations in these genes, which does not account for the fact that passenger mutations are not randomly distributed across the genome and may be recurrently hitting "hot spots" of the genome.

However there are also certain limitations and confounding influences to the evolutionary approach adopted by using the omega ratio. For example, the underlying assumption of this analysis is that since synonymous mutations do not alter the structure and function of proteins they are biologically silent and therefore cannot be selected. However this is now known not to always be the case, as synonymous mutations have been found to frequently act as driver mutations in human cancers [Supek et al., 2014]. It is estimated in Supek et al. [2014] that between one in two and one in five synonymous mutations in oncogenes have been selected and therefore contribute to human cancer. Hence the omega approach may be underestimating the effects of synonymous mutations and confounding results.

Detecting selection in tumour cells cannot only highlight somatically mutated genes and molecular functions that are beneficial to cancer progression by looking for signals of positive selection, but also somatic mutations that are inhibitory to the development of cancer by detecting negative selection in these genes.

1.6.2 Previously used approaches to detect selection in cancer

The omega ratio has previously been used in cancer studies to investigate the relative contributions of driver and passenger mutations in human cancer genomes.

1.6.2.1 Site counts model

Some studies have used a "site counts model", in which a rate of mutations is calculated, e.g. C→T per C nucleotide. Although this model does not intrinsically account for whether the changes are synonymous and non-synonymous, a separate omega ratio can

be calculated at each site by calculating the non-synonymous substitution rate at non-synonymous sites and dividing by the synonymous substitution rate at synonymous sites. [Greenman et al., 2007] have used this approach in the meta-analysis of protein kinase data. A deviation of the omega ratio from that expected by chance (>1) was then interpreted as an indication of the presence of selection upon non-synonymous mutations by the cancer. Over all coding regions of 518 kinase genes, a selection pressure of 1.29 was calculated showing evidence of positive selection by demonstrating an excess of non-synonymous mutations compared with that expected and therefore suggesting the presence of driver mutations in the dataset. Even after the removal of known cancer genes, the selection pressure indicated the presence of driver mutations in genes not previously known to be implicated in cancer.

A limitation with this framework however is that a site counts model does not consider that some nucleotide sites can act as both synonymous and non-synonymous depending on the change that has occurred. This is a slightly crude approach which can be improved by counting every nucleotide three times, once for each possible change, which is how the software suite PAML [Yang, 2007] overcomes this difficulty in Markov models of nucleotide substitution. Although this alters the rate estimates, the rate ratio is preserved.

Greenman et al. [2007] also showed innovative thinking in introducing analysis of mutational profile differences between cancers. For example, they showed that the selection pressure was lower in cancers with defective DNA mismatch repair compared with MMR-proficient cancers. However, since MMR-deficient cancers have a higher prevalence of base substitutions than MMR-proficient cancers, it is possible that driver mutations were overwhelmed by the higher number of passenger mutations in MMR-deficient cancers, suggesting that this model encounters limitations when mutation rates are very high. Greenman et al. [2007] also considered partitioning genes into sub-regions with respect to detecting selection, and grouping by pathway and kinase sub-class. Genes were divided into two groups: kinase domain and everything outside the kinase domain. Over all genes they found a slightly higher selective pressure within

kinase domains, with a much higher selective pressure observed in the P-loops and activation segments of the kinase domains, however these results also suggest that driver mutations are present outside of the kinase domains too. Pathway analysis was carried out by combining Reactome, Panther and INOH in a merged pathway database to test for the presence of mutated pathways, of which the FGF signalling pathway showed the highest enrichment of non-synonymous mutations. After grouping kinase genes by subclass, the highest selection pressure was observed in calmodulin-dependent protein kinases. Previous to this analysis, most reported protein kinase cancer genes were members of the tyrosine kinase or serine/threonine kinase subclasses, so this analysis suggested that other subclasses are also involved in oncogenesis.

1.6.2.2 Codon model

The omega ratio can also be calculated using a codon model, in which the single base changes are not considered directly, instead just the change of one codon to another (e.g. ATG→ ACG) and whether that change is synonymous or not. This overcomes the problem of ambiguity of whether a nucleotide site is synonymous or non-synonymous, as when the whole codon is considered the change can only be either non-synonymous or synonymous and not both. This has been implemented in the more advanced evolutionary style using likelihood models in Yang et al. [2003]. However, their weakness is that they apply their method to just one gene in the genome, examining the spectrum of p53 mutations in cancer. However they do not just use the gene as the unit of analysis, splitting the gene by functional domains, the results of which suggest that amino acid changes less often lead to cancer in structural domains compared with DNA-binding and transactivation domains. Goldman and Yang [1994] also use a codon model which is implemented in the codeml program from the PAML package to deal with single base substitutions only. Whelan and Goldman [2004] have also used a codon model similar to the method used in Goldman and Yang [1994], however they have built on this approach by including di-nucleotide and tri-nucleotide substitutions as well as point mutations. Greenman et al. [2006] have adopted a codon-based evolutionary approach

which has the advantage of being able to accommodate INDELs as well as substitutions, however in doing so this has made it a less robust model in terms of its statistical methodology than others such as [Yang et al., 2003] that just look at point mutations.

Evolutionary codon models already exist for protein-coding data, partly because functionally important variants are more likely to be found in the coding regions of the genome, hence why whole-exomes are preferentially used over whole-genomes in these types of driver detection analyses. Other benefits of exome studies is that there are more cancer exomes currently available than genomes, due to their relatively low sequencing cost. However this is starting to change and as sequencing costs continue to fall, more whole-genome sequences will become available for variant analysis, as well as more advanced evolutionary models to deal with them.

1.7 Personalised medicine

Recent technological advances in genomics are starting to allow for treatments tailored to individuals to become a reality for cancer diagnosis and treatment. There have been many studies dedicated to identifying the mutations involved in the development of cancer, however it is only in the last decade that technological advances in genomic analysis have started to deliver on the promise of personalised medicine [Fernald et al., 2011], gearing towards more targeted treatment options for patients guided by the predictive value of these studies.

Medical sequencing of tumours to guide treatment is one of the first practical realisations of true genome-based personalised medicine [Fernald et al., 2011]. An example of this application currently in action is the use of the chemotherapy medication imatinib, which works by inhibiting the proteins encoded by the ABL and KIT genes that are mutated and activated in chronic myeloid leukemia and gastrointestinal stromal tumours respectively [Gambacorti-Passerini, 2008]. Another example is the use of trastuzumab, which is an antibody directed against the protein encoded by ERBB2 (HER2), which

is commonly amplified and overexpressed in ∼20% of breast cancers [as reviewed in Stratton, 2011, Stratton et al., 2009].

BRAF is also an attractive target for drug discovery strategies. Somatic mutations were discovered in this gene in an early systematic sequencing screen in 2002. This gene encodes a serine-threonine kinase and is mutated in 50-70% of malignant melanomas, 10-15% of colorectal cancers and 50% of papillary thyroid cancers. A substitution mutation at valine 600 with glutamic acid (V600E) accounts for more than 90% of all mutations resulting in constitutive activation of the BRAF encoded kinase. Due to its deep ATP-binding pocket in which inhibitors can sit and because the BRAF V600E mutation is activating, encouraging results have been seen in clinical trials testing inhibitors of V600E mutant BRAF in 80% of patients with malignant melanomas carrying the V600E mutation [as reviewed in Stratton, 2011].

Targeted drug development has also helped to elucidate the mechanisms of drug resistance in recurrent tumours that arise from minor sub-clones containing resistance mutations after drug treatment. Investigations into the genomes of recurrences has identified some of the mutations that confer resistant, which in turn has offered opportunities for new cancer therapies [as reviewed in Stratton, 2011].

Direct targeting and inhibition has predominantly been used to target oncogenes. However this approach becomes a challenge when the mutated gene in question is a tumour suppressor gene, since the proteins encoded by these genes are already inactivated by their mutations. For this, the development of drugs that exhibit synthetic lethality with the mutated cancer gene may be necessary, in which the cancer mutation and the drug together cause the tumour cells death [Garber, 2002]. For example, two genes are synthetic lethal if mutation of either gene alone is compatible with viability but mutation of both leads to death. So inhibiting the product of a gene that is synthetic lethal to the inactivating cancer-causing mutation in a tumour suppressor gene should kill cells that harbour such mutations, while also having the benefit of sparing normal cells [Kaelin, 2005].

Many advances in molecular science have been made in the last decade to benefit medicine, and next generation sequencing has revolutionised the discovery of genes underlying cancer. These efforts include initiatives such as the Human Genome Project [Lander et al., 2001], the International HapMap Project [Gibbs et al., 2003] and the 1000 Genomes Project [Consortium et al., 2012], and now the 100,000 Genomes Project which is currently underway. By combining genetic associations with phenotypes and drug response, personalized medicine has the potential to tailor treatments to a patient's specific genotype [Fernald et al., 2011].

1.8 Aims of investigation

I plan to use evolutionary signatures to uncover evidence of positive selection in cancer exomes, in order to functionally implicate driver mutations important for cancer development, progression and survival.

1.8.1 Specific research objectives

Using well-established techniques from the field of molecular evolution, I wish to answer the following questions:

- Which genes are enriched for driver mutations in cancer? Identify genes harbouring driver mutations (rediscovery of known genes), and uncover candidate driver mutations in genes.
- How does the tissue of origin influence the genes hit by driver mutations in cancer?
- To what extent are mutations in genes dictated by the mutation spectra?
- Is the path of selection and development of a cancer regulated by mutation spectrum or the tissue of origin? Which has more effect on the pathway of mutation? For example, APC is frequently knocked out by premature stop codons in colorectal cancer, caused by C→T changes at CpG sites in mismatch repair, which

suggests that the mutation spectra may be the driving force of these particular mutations in this gene. However, is it this mutation spectra or the tissue of origin (colorectal) that has more of an effect on the mutation pathway of this particular cancer?

Ultimately the aim of identifying key changes in cancer genomes is to further understand how genes are mutated in cancers relating to their mutation spectra and tissue subtypes, to help aid the rational design of target cancer treatments and mediate prognosis.

1.8.2 Approach

In order to achieve these objectives, I propose an alignment strategy in which cancer-specific mutations are edited onto the most up-to-date build of the human reference genome (hg19). I then plan to use omega-based evolutionary analysis on a per-gene basis to detect signals of positive selection in these alignments indicating driver mutation enrichment, in combined as well as sub-type analyses of diverse tumour types. I will also stratify the data based on mutation spectra classifications before evolutionary analysis. Validation will be performed by rediscovery of known cancer genes.

To account for certain limitations of previous cancer studies, the varying coverage in exome sequences will be quantified and accounted for by annotating nucleotide sites as missing in cases where there is not sufficient coverage to confidently call a variant. All exomes will be aligned consistently to hg19 reference genome in order to generate a uniform set of sequences.

Chapter 2

Methodology and data sources

This chapter describes the exome sequences and mutation data used, the data processing pipeline and how the data was prepared for evolutionary analysis. It also comprehensively documents the software versions, parameters and models used for evolutionary analysis.

2.1 Datasets

2.1.1 The Cancer Genome Atlas (TCGA) data

The Cancer Genome Atlas (TCGA) is an ongoing large-scale genome sequencing project [Cancer Genome Atlas Research Network, 2008].

A summary of the data we have used from TCGA for this project is shown in Table 2.1. This dataset consists of paired tumour and normal exomes for 1005 patients over 17 cancer types, generated by next-generation sequencing. Exome sequencing refers to the technique of sequencing all the protein-coding regions of the genome known as the exome (exons of genes), which consists of first capturing all protein-coding exons, and then sequencing the targeted DNA using high-throughput massively-parallel, next-generation sequencing (NGS) technology. The NGS platform used by TCGA to

TABLE 2.1: **TCGA dataset.** This table summarises the paired tumour and normal exome sequence data obtained from TCGA consisting of 17 different cancer types over 1005 patients, with a breakdown of the numbers of patients for each cancer type. *Full TCGA patient IDs and sample types can be found in Supplementary Appendix A*

Cancer type	Patient number
Acute myeloid leukemia (LAML)	53
Bladder urothelial carcinoma (BLCA)	15
Brain lower grade glioma (LGG)	50
Breast invasive carcinoma (BRCA)	110
Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC)	14
Colorectal carcinoma (CRC)*	10
Glioblastoma multiforme (GBM)	208
Head and neck squamous cell carcinoma (HNSC)	85
Kidney renal clear cell carcinoma (KIRC)	175
Kidney renal papillary cell carcinoma (KIRP)	16
Lung adenocarcinoma (LUAD)	26
Lung squamous cell carcinoma (LUSC)	53
Ovarian serous cystadenocarcinoma (OV)	75
Prostate adenocarcinoma (PRAD)	39
Stomach adenocarcinoma (STAD)	19
Thyroid carcinoma (THCA)	19
Uterine corpus endometrial carcinoma (UCEC)	38
Total	**1005**

*Colorectal carcinoma (CRC) is a merged dataset including both colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) patients, however in this case the 10 CRC patients were all COAD.

sequence exomes was the Illumina/Solexa Genome Analyser (Solexa), which is the most widely used next-generation DNA sequencing technology. Other platforms used include 454/Roche Genome Sequencers, SOLiD and Polonator [Shendure and Ji, 2008].

Mitochondrial DNA (mtDNA) is not targeted in currently used exome-sequencing methods [Samuels et al., 2013], however it has been captured in the exome data from TCGA, and so mutations in the mitochondrial genome have also been used in the TCGA analysis.

TABLE 2.2: **Published Lawrence dataset.** This table summarises the cancer-specific mutation data available in the published Lawrence et al. [2014] dataset, consisting of 21 different cancer types (12 from TCGA and 14 from Broad Institute) over 4,728 patients, with a breakdown of the numbers of patients for each cancer type. Cancer types highlighted in red have been used for evolutionary analysis in Chapter 5. These patients contain both coding and non-coding SSNVs.

Cancer type	Patient number
Acute myeloid leukemia (LAML)	196
Bladder (BLCA)	99
Breast (BRCA)	892
Carcinoid (CARC)	54
Chronic lymphocytic leukemia (CLL)	159
Colorectal (CRC)	233
Diffuse large B-cell lymphoma (DLBCL)	57
Esophageal adenocarcinoma (ESO)	141
Glioblastoma multiforme (GBM)	291
Head and neck (HNSC)	384
Kidney clear cell (KIRC)	417
Lung adenocarcinoma (LUAD)	404
Lung squamous cell carcinoma (LUSC)	177
Medulloblastoma (MED)	92
Melanoma (MEL)	118
Multiple myeloma (MM)	206
Neuroblastoma (NB)	76
Ovarian (OV)	316
Prostate (PRAD)	137
Rhabdoid tumor (RHAB)	32
Endometrial (UCEC)	247
Total	**4728**

2.1.2 Lawrence data

Table 2.2 shows a summary of the data we have used from the published work of Lawrence et al. [2014]. This dataset consists of cancer-specific mutations for 4,728 patients over 21 cancer types. They have also used next-generation paired tumour and normal exome sequences. Mutations in mtDNA were not been included in the mutation data provided by [Lawrence et al., 2014].

TABLE 2.3: **Patient overlap between TCGA and Lawrence datsets.** This table summarises the data present in both datasets, consisting of 11 cancer types over 702 patients, with a breakdown of the numbers of patients for each cancer type. *Full patient IDs and sample types of the patients that appear in both datasets can be found in Supplementary Appendix B. TCGA nomenclature has been used in this table, however variations of cancer type names used by Lawrence et al. [2014] can be found in Appendix A*

Cancer type	Patients
Acute myeloid leukemia (LAML)	51
Bladder urothelial carcinoma (BLCA)	15
Breast invasive carcinoma (BRCA)	107
Colorectal carcinoma (CRC)	8
Glioblastoma multiforme (GBM)	171
Head and neck squamous cell carcinoma (HNSC)	85
Kidney renal clear cell carcinoma (KIRC)	150
Lung adenocarcinoma (LUAD)	26
Lung squamous cell carcinoma (LUSC)	50
Ovarian serous cystadenocarcinoma (OV)	4
Uterine corpus endometrial carcinoma (UCEC)	35
Total	**702**

2.1.3 Overlap between TCGA and Lawrence datasets

Table 2.3 shows that 702 patients that overlap between the TCGA samples fully re-processed in this work and those present in the Lawrence et al. [2014] data.

2.2 Obtaining exome sequences and mutation data

2.2.1 TCGA schema parsing

Next-generation whole exome sequences for both primary tumour and matched non-tumour somatic normal samples (usually taken from blood) were obtained from the TCGA Cancer Genome Hub[1]. Paired tumour:normal data was obtained for 1005 patients over 17 cancer types. Sequencing data was downloaded in BAM format, decrypted

[1] https://cghub.ucsc.edu

and validated using Picard version 1.43[2]. Each exome came to ∼20Gb in size, resulting in a 40Tb dataset.

At the initiation of this project the Cancer Genome Hub was not developed and TCGA exome sequences were only available from NCBI dbGaP [Tryka et al., 2014]. Obtaining this data represented a significant barrier to analysis and its use required parsing in order to obtain the samples for analysis. Alignment data could only be searched for, accessed and downloaded based on a SRS ID from dbGaP, which gave no information on the type of data. A schema was developed in order to identify the specific type of data required using Perl scripts. The SRS IDs associated with the data from NCBI were searched for, and then those particular sequences were requested and downloaded using automated FASP downloading of data from NCBI[3]. Samples were subsequently annotated with further information such as their sample type (i.e. tumour or normal) using the sample Atlas ID, and their cancer type (e.g. GBM). Through quality control and data cross-referencing, I identified several miss-annotations and other errors in the dbGaP derived data that were reported back to TCGA for correction.

2.2.2 Retrieving Lawrence mutations

Additional cancer-specific mutations for 4,735 patients over 21 cancer types based on the work of Lawrence et al. [2014] were downloaded as MAF format files from TumorPortal[4] on a per gene basis. However seven of these patients contained no single nucleotide mutations (only INDELs and other non-SNP mutations) so were discarded, resulting in a set of 4,728 patients for analysis. A few mutations were removed from this set of 4,728 patients as they had been classed as SNPs but were in fact longer substitutions, not single nucleotide variants. However this further removal did not alter the number of patients in the set. There is a slight discrepancy between the numbers used in the Lawrence et al. [2014] paper and those available for download. Lawrence et al. [2014] used 4,742 patients in their study, however only 4,735 were available for download.

[2] http://picard.sourceforge.net
[3] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap
[4] http://cancergenome.broadinstitute.org/index.php

Figure S1 produced in the Supplementary material of the Lawrence et al. [2014] study also record using a different number of patients for analysis. 4,729 patients were used to create the figure despite the figure legend stating that 4,742 patients were used.

A Perl one liner was used in a Unix environment to download all available cancer-specific mutations on a per gene basis in MAF format. A list of all gene names was produced from a mySQL database to be passed to the Lawrence website for this retrieval. The bash code for this process is shown in Listing 2.1.

```
# generate list of gene names from mySQL hg19 assembly database
Myblackadder -D codons_hg19 -B -N -e 'select distinct(geneName) from
    genCodons' > geneNamesList
cat geneNamesList | grep -v '\.' > geneNamesList.nodots
cat geneNamesList.nodotsSeps | sort -R > geneNamesList.nodotsSeps.Rand

# perl one liner to download mutation data
for i in cat geneNamesList.nodotsSeps.Rand; do wget -nd http://
    cancergenome.broadinstitute.org/data/per_gene_mafs/$i.maf; sleep 1.$((
    RANDOM%10)); done
```

Listing 2.1: Lawrence mutation data retrieval

2.3 Data processing pipeline

A summary of the data processing pipeline is shown in the flow diagram in Figure 2.1. This mostly involved the exomes obtained from The Cancer Genome Atlas; however, Lawrence et al. [2014] mutations were also incorporated where they could be for consistency. This common pipeline was put in place in order to process this large amount of data efficiently. Unless otherwise stated, the following steps refer to the processing of the TCGA data.

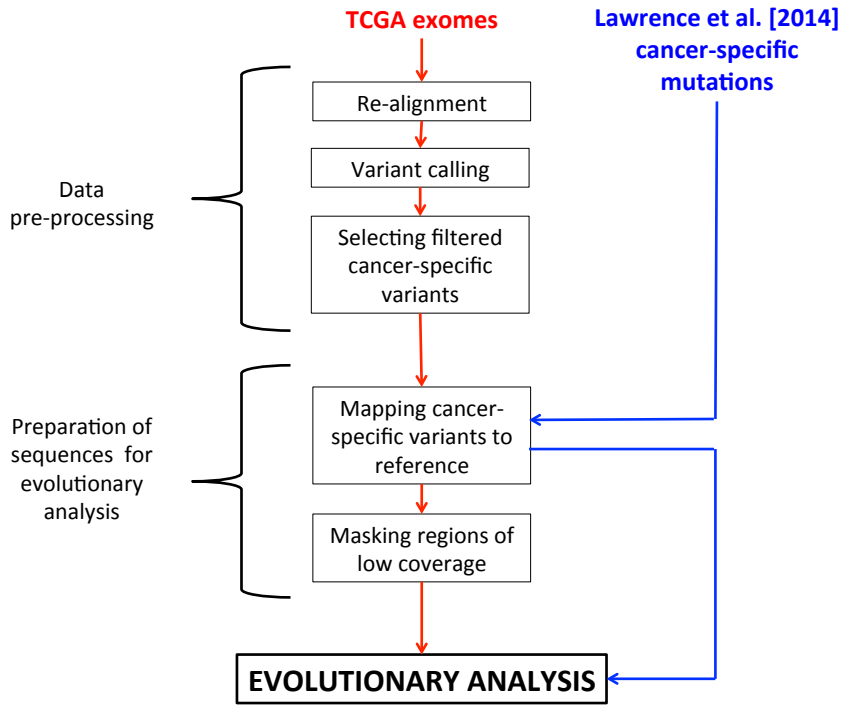


FIGURE 2.1: **Data processing pipeline.** Flow diagram illustrating the methods involved in the data processing pipeline for both the TCGA and the Lawrence datasets prior to evolutionary analysis.

2.3.1 Pre-processing data

TCGA pre-aligned BAM files as obtained were based on a mixture of hg18 and hg19 reference sequence alignments and a heterogeneous combination of alignment tool versions and parameters. These differences were not randomly distributed, but rather they tended to cluster by disease type. Such biases could then lead to systematic biases in a meta-analysis. To avoid such problems, reads were realigned and variants called, selected and filtered in a common pipeline, described here, in the first part of the data processing pipeline.

Initially however, data was handled via two alternative pre-processing approaches:

- **Approach 1:** Variants were called from pre-aligned exomes (BAMs pre-aligned to hg18 reference genome). UCSC liftOver tool [Rosenbloom et al., 2015] was

used to convert the genome coordinates from NCBI build 36 (UCSC hg 18) to the more recent genome assembly NCBI build 37 (UCSC hg 19) positions. This tool uses BED format, so VCF files were converted to BED format first using BEDtools [Quinlan and Hall, 2010].

- **Approach 2:** Improved exome re-alignment to the more recent hg19 version of the human reference genome was carried out prior to variant calling. This approach used state-of-the-art algorithms and a more up-to-date human reference genome to improve exome alignments especially around indels.

The purpose of running these two approaches together was to test the effectiveness of re-aligning all exomes to the most recent reference genome in 'Approach 2', and to ascertain whether this time-consuming process was worth the increased accuracy of called variant results. Therefore, in 'Approach 1' exomes were not re-aligned to hg19 before variants were called, and in 'Approach 2' re-alignment was carried out before variant calling.

Both methods were run simultaneously in order to get quick and easy preliminary results from the faster 'Approach 1', and then eventually higher quality results from the more time-consuming full processing 'Approach 2'. Ultimately 'Approach 2' was adopted for all data pre-processing, since it seemed feasible to be able to process this large amount of data in a reasonable amount of time on the systems available. 'Approach 2' is described in the subsequent re-alignment and variant calling sections.

This pre-processing step was not performed on the Lawrence et al. [2014] dataset, since the cancer-specific variants had previously been called, essentially in the same was as was adopted in 'Approach 1' but with additional filtering applied.

2.3.1.1 Re-alignment to hg19 reference genome

To ensure a directly comparable uniform set of sequences with which to call variants from more accurately, the exomes obtained from TCGA were all processed in the

same way. FASTQ format reads were extracted from BAM files using Picard (version 1.43), and reads were realigned to the most up-to-date GRCh37/ hg19 reference human genome sequence using BWA pre-alignment (version 0.5.9, Li and Durbin [2009]) in conjunction with Stampy (version 1.0.12, Lunter and Goodson [2011]) read mapper, outputting re-aligned reads as BAM files. Aligning to an exome rather than a genome reference would have reduced the computational time, however an exome reference sequence was not available at the time. Additionally, exome reads are often mixed with whole-genome sequencing data due to the nature of the exome capture process before sequencing. Therefore, aligning reads to an exome reference could result in off-target reads being mapped to exons in the reference, which could in turn cause false-positive variant calls. Using a whole-genome reference is hence beneficial in terms of accuracy of downstream variant calling. Alignment processing to sort and merge re-aligned runs was performed with Samtools (version 0.1.14, Li et al. [2009]), a utility suite for handling BAM files. Duplicate reads were marked using Picard version 1.43 [5].

2.3.1.2 Single nucleotide variant calling

Genome Analyis Toolkit (GATK, version 0.5506, McKenna et al. [2010]) was used to generate a set of coordinate intervals identifying indels for local realignment around these gap positions (indels) to improve gap edges. Subsequently, in preparation for improved variant calling, GATK was also used to recalibrate base quality alignment scores based on co-variates calculated across the aligned reads.

GATK (version 0.5974) was then used to uniformly call all single nucleotide variants in tumour:normal pairs (BAM alignments) for each patient using joint calling on the whole genome [DePristo et al., 2011]. At the time exon target regions were not available, so variant calling was not able to be limited to just exons and was instead carried out on the whole genome.

[5] http://picard.sourceforge.net

The tumour exome was compared to both the normal exome and the hg19 reference in order to distinguish the important cancer-specific variants: those present only in the tumour sample; from the germline polymorphisms: present in both the tumour and normal samples but not the reference sequence. Those present in tumour and normal are inherited variants common to all cells of the body (Figure 2.2). There were also cases of control-specific variants called, where the variant was present only in the normal sample. It is possible that this was due to a somatic mutation that had arisen in the control sequence only, however these are rare because the mutation rate in the control sequence is much lower than in the cancer. Another possibility is that it was actually a true germline variant that has been missed in the tumour sequencing. Instances of this could result from a loss of heterozygosity (LOH) event occurring in the tumour so that the copy of the tumour sequence harbouring the germline mutation had been lost and therefore could not be seen. Whichever the case, these mutations are not cancer-specific so are not expected to be important in oncogenesis.

The process of variant calling involved joint calling to get a joint estimate of cancer and control variants, in which the paired tumour and normal samples from the same individual were compared simultaneously to the reference genome for each patient, rather than using single calling in which the control and cancer would be separately compared to the reference and then those results compared. This approach reduced the chance of false-positives (both cancer-specific and control-specific) when germline events were occurring, as more reads were available in the tumour and normal to base a variant call on, which in turn increased the power of the analysis in correctly calling true germline variants. Another type of cancer-specific false-positive occurs when the tumour and normal are both wild-type and an event is inaccurately detected in the tumour. This is likely to be a result of a machine-sequencing error, incorrect local alignment of individual reads and discordant alignment of pairs rather than insufficient coverage. However, paired calling can also help reduce this type of false-positive [Meyerson et al., 2010]. The increased coverage in joint calling also reduced the chance of having to account for missing data, which is dealt with later on in this pipeline. Joint calling

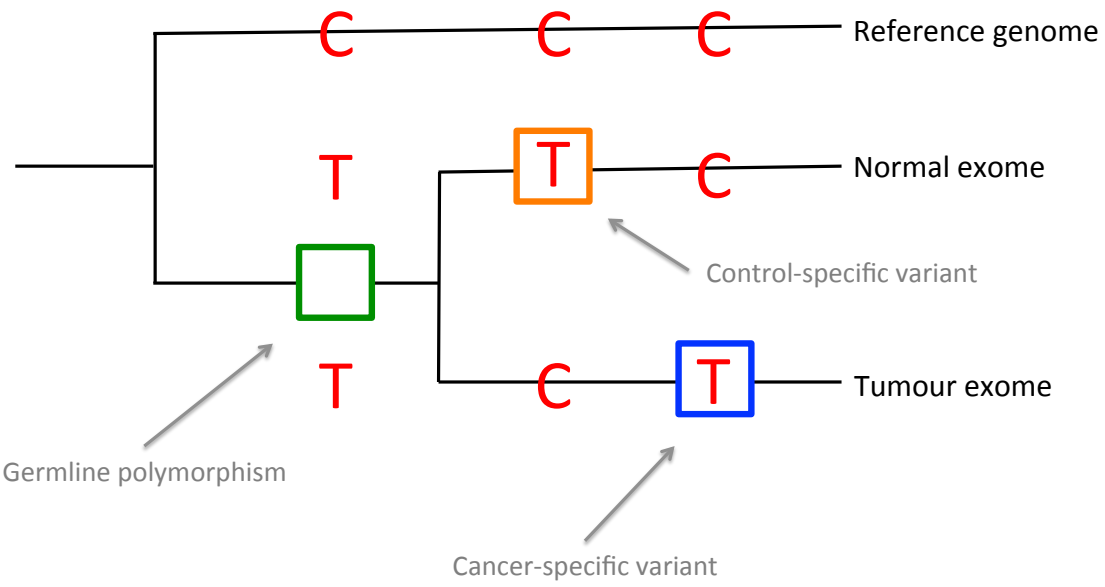


FIGURE 2.2: **Joint single nucleotide variant calling on TCGA data.** Evolutionary tree of the reference, tumour and normal sequences. During joint calling, both the tumour and normal exome sequences were simultaneously compared to the hg19 reference genome sequence in order to identify and distinguish the cancer-specific SNVs from the germline and control-specific SNVs. The green square represents a germline polymorphism, the orange square displays a control-specific variant and the blue square highlights a cancer-specific variant. The red letters denote nucleotide bases in the DNA sequences.

improved the sensitivity and specificity of the variant calling process, making this a more informative way to call SNVs as opposed to single-sample calling.

Nucleotide substitution and small insertion/ deletion (indel) variants were annotated with predicted coding effects using SnpEff [Cingolani et al., 2012], e.g. synonymous vs. non-synonymous. This produced VCF files. The variants were also cross-referenced with dbSNP to show if the variants were known to be segregating in the population or whether they were novel SNVs.

All called variants and consequence annotations were loaded into a MySQL (version 5.5.12) database for processing and interrogation. The schema for the TCGA database in MySQL (called *tcga_pair_exome*) is shown in Figure 2.3.

The main tables in the mySQL database are:

- **sample** - patient/sample information.
- **var_site** - variant site information, containing 107,637,070 different loci with variant sites in at least one sample (many sites are variant in more than one sample).
- **var_site_sample** - cancer/control genotypes and qualities for a given patient at a given site, with 850,586,532 different genotypes (mappings between the variant site and the sample).
- **consequence** - variant consequences, with consequences for each possible transcript that a specific mutation can occur on, with 9,090,332 different protein-coding consequences for the variant sites. Some sites do not have protein-coding consequences (e.g. introns), so will not be included in this table. Others have multiple consequences because the gene in which they are located has multiple transcripts, and the consequence system is based on transcripts rather than genes. This table shows how a mutation affects every transcript in which it occurs. For one transcript for a given gene, the mutation might cause a stop codon but for another it might fall within an intron and have no effect (and therefore will not be in this table).

sample, **var_site** and **var_site_sample** are linked by *sample_id* and *var_site_id* fields. **consequence** is linked to the relevant variant sites through the *var_site_id* field.

The helper tables are:

- **gene_transcript** - maps Ensembl gene and transcript IDs to each other for loading variant annotations from snpEff
- **tss_disease** - maps codes from the atlas IDs to the disease types for each sample

The database contains 1007 patients, however two of these do not contain any cancer-specific SNVs so were not used for any analysis.

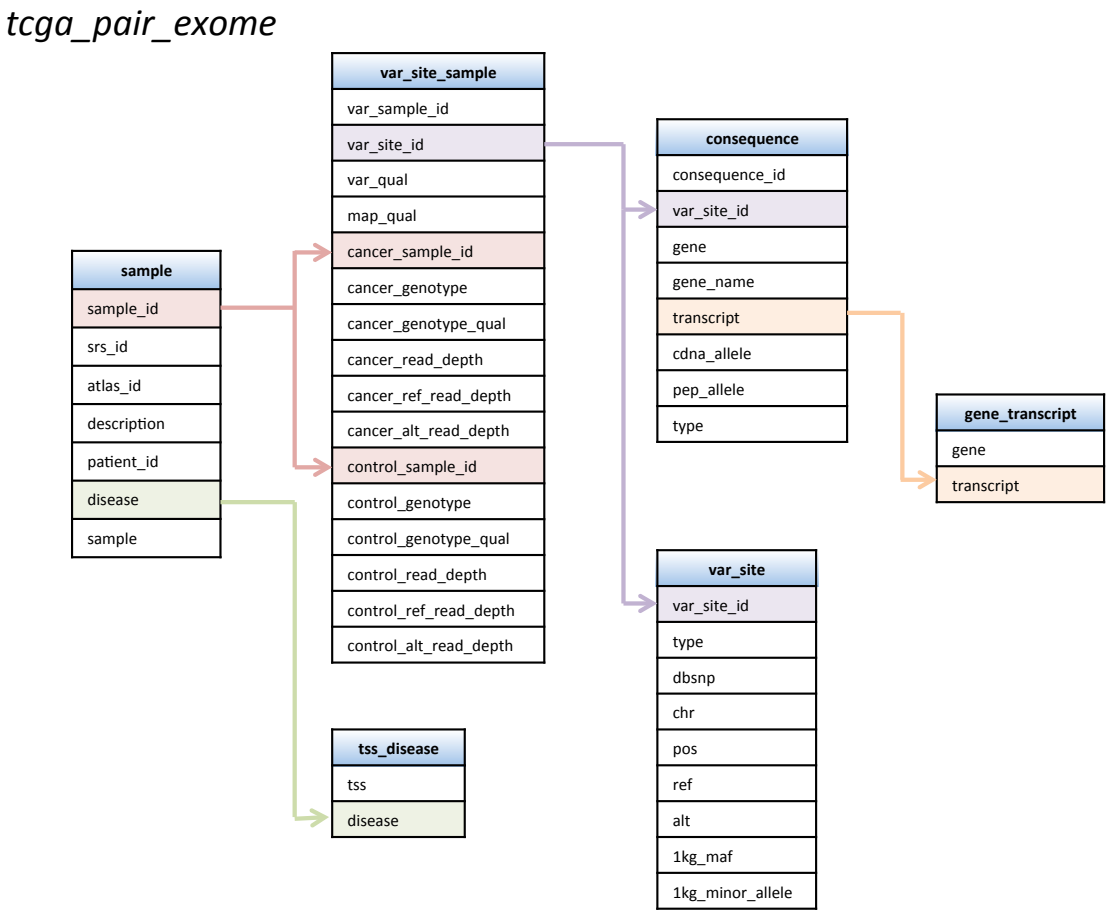


FIGURE 2.3: **MySQL TCGA database schema.** This diagram shows the structure of the MySQL database, *tcga_pair_exome*, containing the TCGA data. This database consists of six tables: "sample", "var_site_sample", "tss_disease", "consequence", "var_site" and "gene_transcript". For each table the fields have been listed in this schema. Arrows show how the tables are linked by fields during database querying.

2.3.1.3 Selecting cancer-specific filtered variants

Using the MySQL database of all uploaded variants called on the 1005 TCGA patients, non-synonymous and synonymous SNVs were selected based on the following criteria:

- **Cancer-specific heterozygous SNVs:** SNVs with a heterozygous genotype (0/1) in the tumour sample and reference genotype (0/0) in the normal sample for a given patient.

- **Control-specific heterozygous SNVs:** SNVs with a heterozygous genotype (0/1) in the normal sample and reference genotype (0/0) in the tumour sample for a given patient.
- **Germline SNVs:** SNVs with either a heterozygous (0/1) or homozygous (1/1) genotype in the tumour and normal sample for a given patient (not necessarily the same genotype in both).

For each of these three categories, a minimum genotype quality of 40 and a minimum SNV site quality score of 40 was used to filter out low quality SNVs. This selection process was carried out purely using Perl scripts to mine the MySQL database of variants.

Stop codon mutations were also re-annotated as non-synonymous or synonymous, so that stop codon mutations were included in the evolutionary analysis. Non-synonymous-start, start-lost, stop-gained and stop-lost mutations were all re-annotated as non-synonymous-coding mutations. Synonymous-start and synonymous-stop mutations were re-annotated as synonymous-coding mutations.

The set of cancer-specific heterozygous SNVs were filtered further by removing those that occurred as control-specific heterozygous SNVs or germline SNVs in other samples, on a per patient basis, to enrich for somatic only mutations. This filtered set of cancer-specific SNVs was used for the next stage of data processing in the TCGA pipeline. In this quality control (QC) step, the cancer-specific heterozygous SNV set prior to filtering consisted of 525,675 variants. However, post-filtering, 244,634 cancer-specific heterozygous SNVs remained, resulting in a loss of 53.46% (281,041) of the original variants.

Homozygous cancer-specific SNVs were not selected in these subsets as they are harder to interpret. They could be errors since they are much rarer than heterozygous SNVs, as it is unlikely that a random somatic mutation will occur in exactly the same position on both alleles (copies) of a gene. Alternatively a germline mutation followed by an additional somatic mutation in the tumour on the other allele could have taken

place (0/1 in control, 1/1 in tumour). However, the model would become much more complicated when having to deal with a genotype of 0/1 in the normal and 1/1 in the tumour. This analysis has therefore been restricted to the classic model of 0/0 in the normal and 0/1 in the tumour to identify cancer-specific heterozygous variants. It is possible that within those homozygous cancer-specific mutations (genotype of 1/1 in the tumour) that have been excluded in this analysis, some could be the result of a loss of heterozygosity (LOH) in the tumour plus a somatic inactivating mutation on the remaining allele, which would appear as homozygous (1/1) in the tumour but is actually representing a heterozygous somatic mutation in the cancer. By filtering out homozygous cancer-specific mutations these types of LOH events are not considered. However, LOH events would be included in the identified germline variants, using the criteria of either heterozygous (0/1) or homozygous (0/0) in the tumour and control samples. With additional sequencing depth information these events could be confirmed (by observing regions of low sequencing depth).

Lawrence et al. [2014] mutations had already been filtered prior to this analysis, so this assignment of cancer-specific variants step was not necessary for the Lawrence data.

2.3.1.4 Ambiguity of SNV and INDEL counting

There was ambiguity in the counting of unique SNVs and INDELs when using the 'consequence' table from the *tcga_pair_exome* mySQL database (containing all filtered heterozygous cancer-specific TCGA SNVs and INDELs). This specific table included, for each unique mutation, the consequence the mutation could have in each possible transcript for that gene. This meant that the same unique mutation in a gene in a patient was sometimes represented multiple times, once for each possible transcript that the mutation could occur on for that gene for any given patient. Therefore it was important to consider this and ensure that each unique mutation was counted only once regardless of how many transcripts existed for that gene.

To create a unique set of SNVs for gene/patient/disease/genome-based analysis in Chapter 3 and deal with the ambiguity of counting unique mutations multiple times, for each mutation in a gene for any given patient the transcript with the single most severe consequence was used for counting annotation types. Since some genes in the genome are overlapping, for cases where a single mutation affected two genes, the mutation was counted twice (in both genes) for gene-based analysis, but only once for patient/disease/genome-based analysis.

This same system was used for INDEL counts in Chapter 7.

To resolve this issue in the evolutionary analysis in Chapters 4, 5 and 6, in overlapping genes the SNV mutation was counted in both genes, so the mutation was counted twice, and the consequence from the mutation occurring on the longest transcript was used before being edited onto the longest reference transcript.

2.3.1.5 Data management

In total the data pre-processing, consisting of re-alignment and variant calling, took approximately ten days to process a tumour and normal sequence for a single patient, the re-alignment being the time-limiting step. Without the re-alignment (as described previously in "Approach 1"), this pipeline only took ∼36 hours to run per patient.

However, the running time of "Approach 2", that was adopted in this analysis, was optimised by processing multiple samples in parallel. Figure 2.4 shows the rate at which TCGA data was becoming available compared to how fast it was able to be processed. The black curve shows the cumulative availability of TCGA data, and the red and green curves represent the rate at which the data was processed using "Approach 1" and "Approach 2" respectively, as described earlier. The two straight lines show the predicted rate of analysis that could be achieved for "Approach 2". The purple line was the expected rate if 32 exomes were to be run simultaneously, and the blue straight line demonstrated the predicted rate at which 48 exomes could be run simultaneously. As

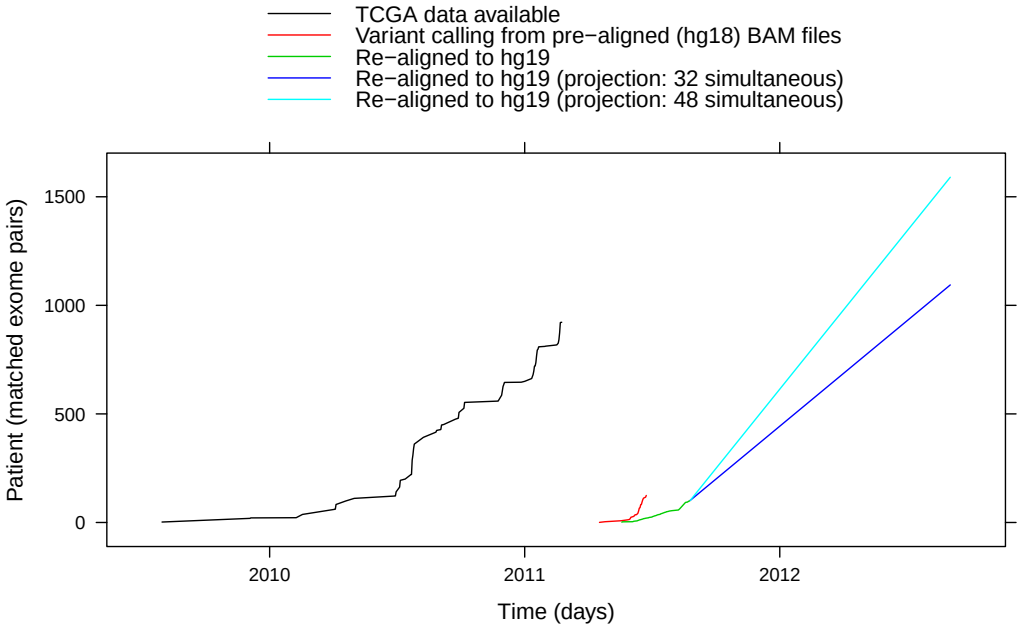


FIGURE 2.4: **Measuring and estimating computational run times for TCGA data.** This plot shows the time scale of running both "Approach 1" (in red) and "Approach 2" (in green). A projection of expected rates using "Approach 2" is shown by the straight lines for running 32 (purple) and 48 exomes (blue) in parallel.

can be seen from the graph, running 48 exomes in parallel was predicted to be sufficient in order to keep up with the rate of incoming data.

2.3.2 Preparation of sequences for evolutionary analysis

Prior to evolutionary analysis, sequences were produced by editing the filtered set of cancer-specific TCGA variants, as well as the cancer-specific Lawrence mutations, onto reference transcripts. For TCGA data only, regions of the consensus sequence were annotated as missing if coverage was too low to confidently call a variant. Finally, edited transcript sequences were converted into the correct format for evolutionary analysis, with one file per gene containing transcripts aligned over all patients for that gene.

2.3.2.1 Editing TCGA and Lawrence cancer-specific variants onto reference sequences

To ensure only cancer-specific SNVs were present in the sequences used for evolutionary analysis, reference sequences were used and just the cancer-specific variants were edited onto these sequences. The reason for preparing new sequences and not using the tumour sequences already available, was to ensure that the germline polymorphisms which were originally present in the tumour sequences were not present in these new sequences and therefore could not confound the evolutionary analysis (which was used specifically to test for positive selection acting on somatic driver mutations in cancer exomes).

The longest reference transcript sequences for each gene, together with the longest transcript exon coordinates and corresponding Ensembl gene IDs, were obtained from Ensembl API version 65 for the TCGA data and Ensembl version 75 for the Lawrence sequences [Flicek et al., 2014], both Ensembl versions based on the hg19 reference genome assembly. Transcript IDs change with new Ensembl versions (new annotations), so the longest transcript for each gene may differ between the TCGA and Lawrence methods.

The filtered set of non-synonymous and synonymous heterozygous cancer-specific TCGA mutations were mapped and edited onto the longest hg19 reference transcripts (coding sequence) for the relevant protein-coding gene from Ensembl 65 using the Ensembl API and Perl (version 5.14.1)[6]. The set of pre-called filtered Lawrence et al. [2014] cancer-specific mutations were also edited onto the reference transcripts in the same way, but using longest transcripts from the more recent Ensembl version 75 instead. For each patient, only one transcript sequence was used for editing. This process is shown in Figure 2.5.

For the Lawrence variant data only, the following information was first required before editing, since the TCGA variants were annotated with this information in the previous variant calling stage:

[6] http://www.perl.org/

```
Patient 1:  ATGGTCTCCACCTACCGGGTGGCCGTGCTGGGGGTG...
Patient 2:  ATGGTATCCACCTACCGGGTGGCCGTGCTGGGGGCG...
Patient 3:  ATGGTCTCCACCTACCGGGTGGCCGTGCTGGGGGTG...
Patient 4:  ATGGTCTCCACCTACCGGGTGGGCGTGCTGGGGGCG...
Patient 5:  ATGGTCTCCACATACCGGGTGGCCGTGCTGGGGGCG...
Patient 6:  ATGGTCTCCACCTACTGGGTGGCCGTGCTGGGGGTG...
Patient 7:  ATGGTCTCCACCTACTGGGTGGCCGTGCTGGGGGCG...
Patient 8:  ATGGTGTCAACCTACCGGGTGGCCGTGGTGGGGGCG...
```

FIGURE 2.5: **Edited TCGA and Lawrence reference transcripts.** Each line represents the DNA sequence for a single patient in the alignment, as only one transcript was used per patient. Cancer-specific SNVs edited onto both the TCGA and Lawrence reference transcripts are highlighted in red.

- Ensembl gene ID
- Ensembl transcript ID
- Peptide position

The Ensembl gene ID was necessary to map the mutation to the longest reference transcript sequence obtained from Ensembl, and the peptide position differs depending on the transcript the mutation has occurred in, so only the peptide position with the longest Ensembl transcript ID was used for editing the mutation onto the longest reference sequence.

This information was obtained from Ensembl (v75) using Variant Effect Predictor (VEP) [McLaren et al., 2010], which requires the following input information for the Lawrence variants:

- Genomic position
- Chromosome
- Reference allele
- Newbase

The Lawrence cancer-specific mutations were annotated with this additional information before they were able to be edited onto the longest reference transcripts.

The mutation information input into VEP does not change with Ensembl version, not unless the assembly changes which it has not in this case. However, the output annotations from VEP can change with new Ensembl versions, therefore the Lawrence data will have more up-to-date Ensembl (v75) annotations than the TCGA data which was processed using Ensembl 65.

At this point of data processing, some SNVs were not edited onto the reference sequences if the variant did not occur in the longest transcript for that gene. SNVs were also excluded if more than one mutation occurred in the same codon, as this rare occurrence complicates some of the assumptions underlying the analyses used later.

The sequences edited with the Lawrence cancer-specific SNVs were then ready for evolutionary analysis.

2.3.2.2 Missing data annotation

During DNA sequencing, the depth of coverage (sequence depth) varies across the genome. For example, CpG regions are particularly prone to low coverage depth, partly due to these CG-rich regions remaining annealed during amplification in sequencing [Veal et al., 2012]. The coverage depth refers to the number of times a nucleotide is read during sequencing [Sims et al., 2014], so the coverage of reads over a site in a sample is the alternate read depth plus the reference read depth. Sequence depth is measured by the amount of over-sampling. In whole genome sequencing, to detect mutations with high sensitivity the 3 billion nucleotides of the human genome are covered at least 30-fold, which requires the generation of 90 billion bases of sequence data per sample. However, in cancer samples the number must be increased to account for the decreased purity and increased ploidy[Meyerson et al., 2010]. The lower the coverage, the harder it is to correctly genotype a site from next-generation sequencing data.

Coverage data has been calculated for the TCGA data, so to account and correct for the problem of varying coverage, regions of low coverage were annotated as missing on the TCGA transcript sequences previously edited with cancer-specific variants. Where there was insufficient coverage in either the tumour or control sample to be able to confidently call a variant, the position was annotated as missing. This process was put in place to help reduce the rate of calling cancer-specific false-negatives where coverage was poor in the tumour, as well as cancer-specific false-positive mutations where coverage was reduced in the normal. A false-positive could result from a true heterozygote call for a non-reference allele (true germline mutation) present in the normal sample that was missed due to a lack of read coverage at that site in the normal. Conversely, a false-negative would result from low coverage in the tumour sample, at sites where a true cancer-specific variant is present but is not detected due to low coverage depth in the tumour. This was an important step in ensuring that cancer-specific variants were not underestimated in cases where there was a lack of coverage in the tumour sequence, resulting in cancer-specific false-negatives which would lead to under-representation of selection in subsequent evolutionary analysis. This step was also necessary in avoiding the over-estimation of variants in cases where there was a lack of coverage in the normal sequence, resulting in control-specific false-positives and potentially causing true germline mutations to appear as cancer-specific variants.

Germline variants could be mistaken for control-specific variants in the same way, by missing variants present in the tumour sample, resulting in a higher than expected rate of control-specific SNVs due to false-positives. Control-specific false-positives could also be caused by MMR-deficient cancers or drug treated cancers in which the incidence of LOH is increased (loss of allele in the tumour sample) so that again true germline polymorphisms appear as control-specific. However, much less control-specific variants are expected compared to cancer-specific variants, due to the mutator phenotype in cancer.

A read depth threshold for missing data was chosen using the empirically calibrated SNV detection sensitivity correction plot in Figure 2.6, which was developed to quantify

how much variation is missed at a given coverage [Meynert et al., 2014, 2013] for heterozygous SNV sites. This calibration recall curve sensitivity was created using germline heterozygous coding SNVs from TCGA control (non-tumour) data, for whole exome sequences (plotted in pink) and whole genome sequences (plotted in blue) for comparison using the same samples, and from heterozygous HapMap 3.3 [Consortium et al., 2010] variant positions located within coding sequence as determined by Ensembl 72 [Flicek et al., 2013] as a gold-standard set of SNP calls for each sample [Meynert et al., 2014]. Since this project used whole exome sequences, the relevant curve is the pink TCGA-WXS curve. Along the x-axis is the read depth, and along the y-axis the sensitivity at varying coverages is shown. The sensitivity measures the confidence that a heterozygous SNV site will be detected if it is truly present. For example, a heterozygous site with a read depth of 10X corresponds to a sensitivity of 0.95 in whole exome sequences (shown by black horizontal line on plot), estimating that there is a 95% chance that a true-positive heterozygous SNV will be called at a site with a coverage of at least 10X in normal samples. These heterozygous sites (with a coverage of >=10X) in the normal sample therefore have a <5% chance of missing a heterozygous SNV in the normal (and calling a false-positive cancer-specific variant which is actually a germline polymorphism).

A read coverage filter of 10X was chosen as the threshold in this analysis, and was applied to the TCGA normal samples. This coverage cut-off was chosen as it has previously been shown in Meynert et al. [2014] that this threshold of 10X at per-site mapped depths results in a 95% sensitivity in detecting heterozygous SNPs in TCGA whole exome sequences. Therefore, since only the heterozygous cancer-specific SNVs were edited onto the reference transcript sequences in the previous step of this analysis, by masking all sites with a read depth of <10X all other polymorphic sites can be reliably called at a sensitivity of 95%. For consistency the same read coverage filter was also applied to the tumour samples (to avoid false-negative cancer-specific variants).

However, the problem of cellular heterogeneity encountered in cancer populations makes this recall ability correction slightly less reliable in tumour samples, as it can not be

known precisely the likelihood of missing a variant that is present in the tumour sample. However, it is especially important to account for this missing variant calls bias in cancer data compared to other types of data because of the often low frequency at which cancer mutations can be found in a heterozygous cell population. So although tumour heterogeneity makes this correction method difficult in terms of reliability, it is also more useful in cancer genomes and has not been addressed previously in cancer studies.

Sites with a coverage of <=10X in either the tumour or normal samples were annotated with "N" in the edited sequence alignments using Perl (version 5.14.1).

The transcript sequences edited with cancer-specific mutations were then ready for evolutionary analysis after the missing data had also been annotated onto the sequences, as can be seen in Figure 2.7.

Coverage information was not available for the Lawrence et al. [2014] data and these detection sensitivity issues were ignored in the published analysis of this data, so this step was only carried out on the fully re-processed TCGA dataset.

2.3.2.3 Coverage depth across tumour and normal exomes

The process of targeted capture and sequencing of the exome (exome-seq), as was used for this project, is known to produce a relatively heterogeneous profile of read coverage over target regions when compared to the more homogeneous whole-genome sequencing (WGS). As a result, WGS yields improved SNP detection sensitivity across regions of interest [Meynert et al., 2014]. Site level SNP detection sensitivity is defined as the mapped read depth directly over a polymorphic site that is required to reliably call that polymorphism [Meynert et al., 2014]. Despite this, the exome contains fewer repetitive elements than non-coding regions present in the whole genome, and also contains most of the causal disease variants identified to date, which is why protein-coding exomes continue to be the focus of investigations into disease-causing variants.

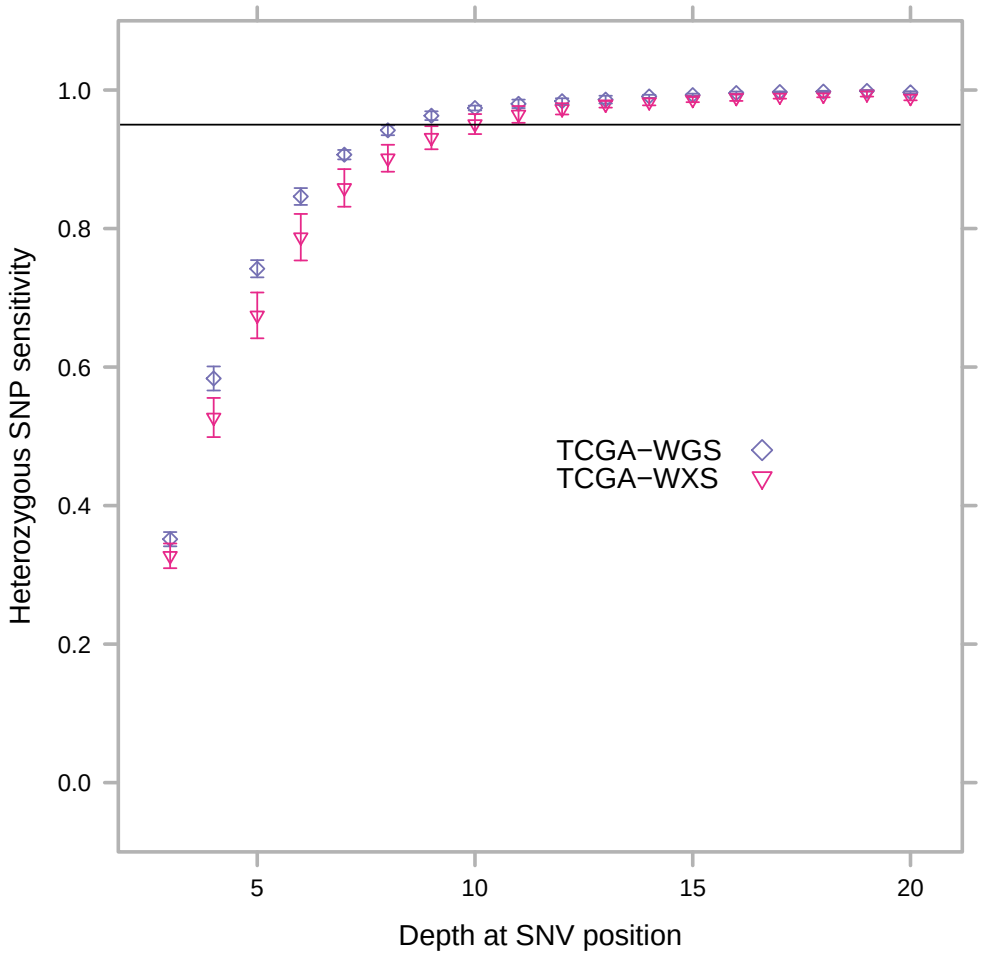


FIGURE 2.6: **TCGA heterozygous SNV detection sensitivity calibration curve.** Plot quantifying how much variation is missed at a given coverage, with depth at heterozygous SNV position plotted along the x-axis, and sensitivity (confidence of calling a true heterozygous SNV if truly present at that position) along y-axis. This plot was generated based on germline heterozygous coding SNPs in whole exome (TCGA-WXS) and whole genome (TCGA-WGS) TCGA control samples. The error bars are one standard deviation from the mean. A horizontal line has been drawn where a depth of 10X meets the TCGA-WXS curve (pink), at a sensitivity of 95% for heterozygous SNVs. *Plot taken from work by Meynert et al. [2014, 2013]*

```
Patient 1:  ATGGTCTCCACCTACCGGGTGGCCGTGCTGGGGGTG...
Patient 2:  ATGGTATCCACCTACCGGGTGGCCGTGCTGGGGGCG...
Patient 3:  ATGGTCTCCNNNNNNCGGGTGGCCGTGCTGGGGGTG...
Patient 4:  ATGGTCTCCACCTACCGGGTGGGCGTGCTGGGGGCG...
Patient 5:  ATGGTCTCCACATACCGGGTGGCCGTGCTGGGGGCG...
Patient 6:  ATGGTCTCCACCTACTGGGTGGCCGTGCTGGGGGTG...
Patient 7:  ATGGTCTCCACCTACTGGGTGGCCGTGCTGGGGGCG...
Patient 8:  ATGGTGTCAACCTACCGGGTGGCCGTGGTGGGGGCG...
```

FIGURE 2.7: **Edited and annotated TCGA reference transcripts.** Reference transcripts edited with cancer-specific TCGA heterozygous SNVs, after having also been annotated with missing data, shown as green "Ns". The red letters denote the cancer-specific SNVs that have previously been edited onto the sequences. Each line represents a different patient, as only one transcript sequence was used per patient.

The issues that negatively impact on SNP detection sensitivity in exome-seq include: PCR amplification, which tends towards lower coverage of GC-rich regions caused by annealing during amplification and compromises the uniformity of coverage; and the preferential capture of reference sequence alleles by exome-seq target probes, which biases the allele distribution away from alternate alleles at heterozygous sites [Meynert et al., 2014]. To account for the reference allele bias in exome-seq, more reads are required to successfully genotype heterozygous SNPs than would be needed in WGS, and to account for the greater variability in coverage in exome-seq a greater mean on-target depth is required to identify the same proportion of SNPs in exome-seq compared to WGS. It is therefore important to access the coverage depth across the exome (Figures 2.8 and 2.9), and quantify the depth required for reliable heterozygous SNP detection in exome-seq, as has been done in this analysis (described above).

Figures 2.8 and 2.9 show the range of coverage across all 1,005 tumour exomes and all 1,005 normal exomes respectively, using a cumulative density function to assess what fraction of the genome is covered by more than a certain number of reads. Figure 2.8 shows that the majority of sites (ranging from 55% to 90% of the target region covered) have $\geq$10X fold coverage in the tumour exomes. Similarly, in Figure 2.9, the majority

of sites (ranging from 60% to 90% of the target region covered) are shown to have ≥10X fold coverage in the normal exomes.

BEDtools (coverage) [Quinlan and Hall, 2010] was used to read in both the 1,005 tumour BAM files and a BED file containing the target capture regions from Illumina, who use Nextera Rapid Capture Exome kit to prepare libraries for exome sequencing. Perl was used to run this function over all 1,005 exomes in batches of 100 exomes, as is shown in Listing 2.2 for 100 tumour exomes. The *-hist* option was used to output a summary histogram for all features in the BED files, which was then used to plot a cumulative frequency distribution in R. This process was repeated for the analysis of coverage in the 1,005 normal exomes.

```
#!/usr/bin/perl -w

open(LIST, "/array11/TCGA/bam_realigned/NEW_TUMOUR/
    BAMfileListTumour_batch1");

my @infiles;

while(defined(my $list=<LIST>))
{
    chomp $list;

    push (@infiles, $list);
}

foreach my $infile (@infiles)
{
    chomp $infile;

    my $cmd = "bedtools coverage -hist -abam /array11/TCGA/bam_realigned/
    $infile -b nexterarapidcapture_exome_targetedregions_v1.2.bed | grep ^
    all > $infile.hist.all.txt";

    print "$cmd\n";
    system ($cmd);
```

```
22 }
```

LISTING 2.2: perl script to retrieve coverage information over target exome regions (run over BAM files in batches of 100)

It can be seen that the coverage varies over each exome in both distributions. It is expected that the uniformity of read coverage would be improved in WGS, as well as reducing bias of allele ratios, both of which would require a lower read coverage than is required for the detection of SNPs in exome-seq. Other benefits of using WGS over exome-seq include the improved detection of genomic rearrangements and disease-causing polymorphisms in non-coding regions of the genome [Meynert et al., 2014].

2.4 Data analysis and software used

2.4.1 Evolutionary analysis in PAML

2.4.1.1 Selection and substitution models used

Evolutionary analysis was carried out in the software suite PAML (Phylogenetic Analysis by Maximum Likelihood) version 4 [Yang, 2007], a package of programs for phylogenetic analyses of DNA and protein sequences using maximum likelihood. The PAML package includes the programs baseml and codeml amongst others. The program baseml is used for the maximum likelihood analysis of nucleotide sequences. The program codeml is formed by merging two old programs: codonml, which implements the codon substitution model of Goldman and Yang [1994] for protein-coding DNA sequences, and aaml, which implements models for amino acid sequences. These two are now distinguished by the variable *seqtype* in the control file *codeml.ctl*, with a "1" for codon sequences and "2" for amino acid sequences. The programs baseml, codonml and aaml use similar algorithms to fit models by maximum likelihood, the main difference being that the unit of evolution in the Markov model, referred to as a "site" in the sequence, is a nucleotide, a codon or an amino acid for the three programs respectively. Markov

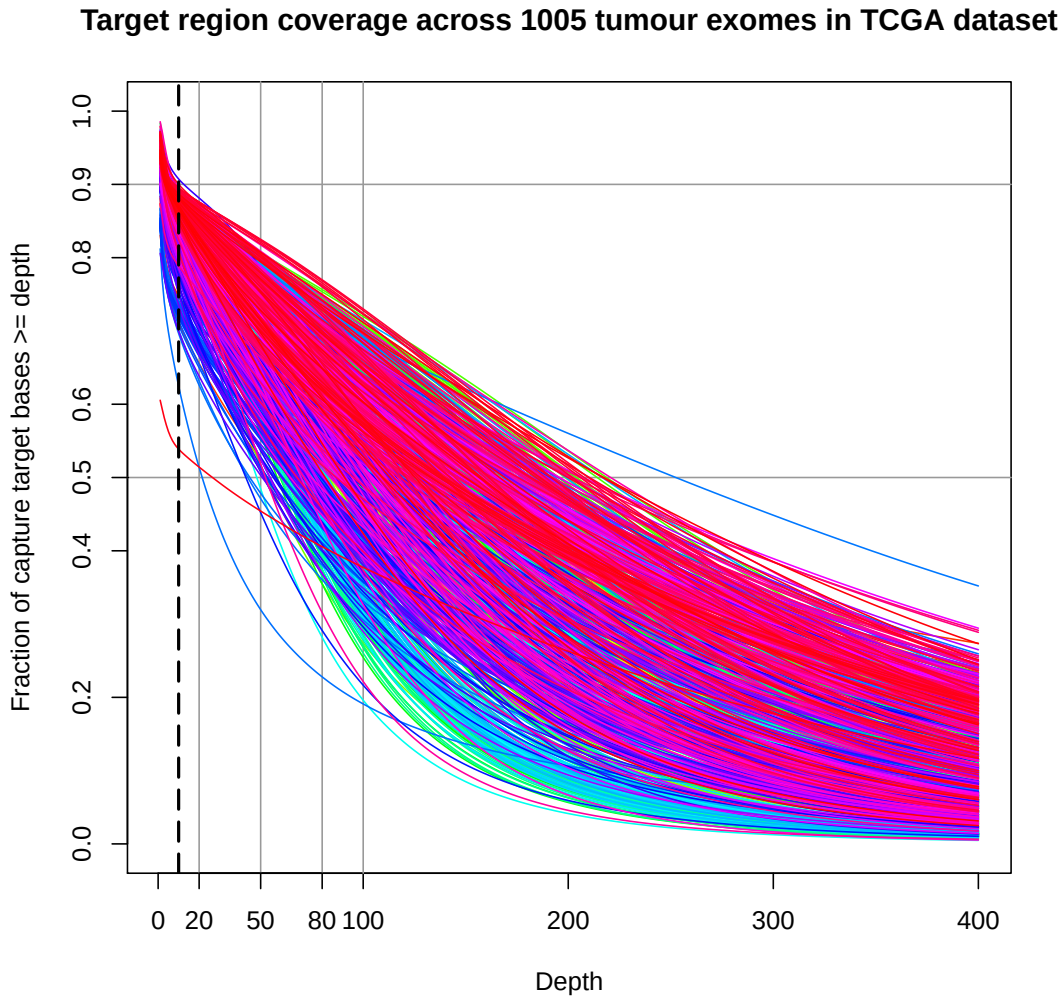


FIGURE 2.8: **Visualisation of coverage depth over 1005 TCGA tumour exomes.** The coverage per base for the normal samples has been plotted as a cumulative distribution describing the fraction of targeted bases that were covered by more than a certain number of reads. The black dashed line denotes a read depth of 10X, which shows that the fraction of capture target bases covered by $\geq$ 10X ranges from 60% to 90% in the normal exomes in the TCGA dataset, with the exception of one exome that appears to have some problems with only 60% of capture target bases with read depth recorded.

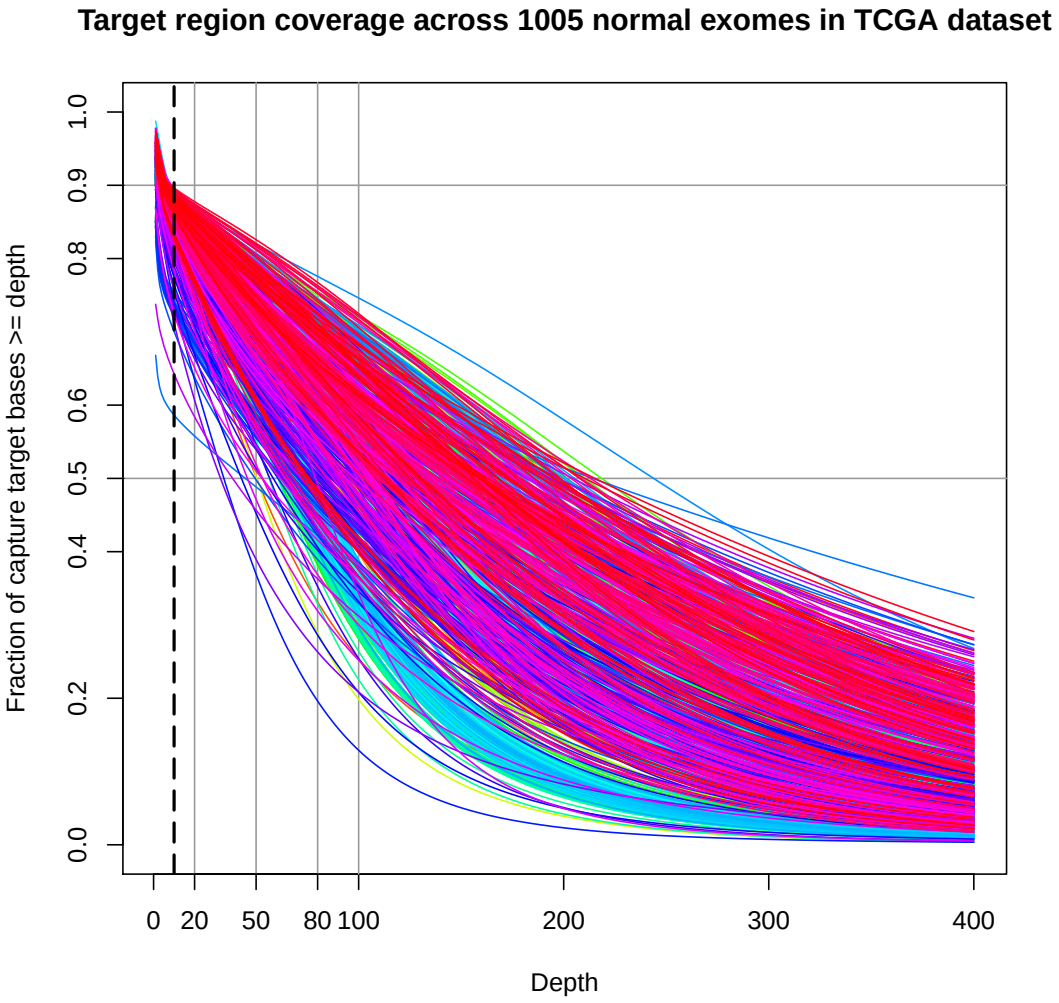


FIGURE 2.9: **Visualisation of coverage depth over 1005 TCGA normal exomes.** The coverage per base for the normal samples has been plotted as a cumulative distribution describing the fraction of targeted bases that were covered by more than a certain number of reads. The black dashed line denotes a read depth of 10X, which shows that the fraction of capture target bases covered by $\geq$ 10X ranges from 60% to 90% in the normal exomes in the TCGA dataset.

process models are used to describe substitutions between nucleotides, codons or amino acids, with substitution rates assumed to be either constant or variable among sites. A sites model was used in this analysis, as described below, which allows the substitution rates to vary among sites.

Using the codeml program within this package, a codon-based model was implemented to analyse single base substitutions and obtain an omega ratio. This model was chosen over a single substitution site model such as JC69 [Jukes and Cantor, 1969], due to the difficulty of dealing with the ambiguity of whether a single site is non-synonymous or synonymous as without considering the sequence context in the codon a single site could be both synonymous and non-synonymous depending on the change. This ambiguity is removed at the codon level, since the codon site can only be non-synonymous or synonymous. Therefore the added context of the codon sequence makes a codon model more robust than a single nucleotide site model.

Parameters were estimated that provide the model that best describes the observed data, using the **codeml** program within this package, to implement a codon-based model which . Analysis in PAML was performed on a per gene basis. Three files for each gene were required for the running of PAML: a control file (*codeml.ctl*), a tree structure file (*codeml.dnd*) and a sequence data file (*codeml.aln*).

The tree file contained a list of the patient IDs in the sequence data file. This file shows how the sequences are related, which in this case is equally as they are artificially generated as opposed to species' sequences.

The sequence data file for each gene contained the sequence alignments prepared for PAML, over all patients in that gene with one reference transcript sequence per patient containing edited cancer-specific mutations. The "native" format in PAML for this file is the PHYLIP format [Felsenstein, 2005] shown in Figure 2.10. The first line contains the number of sequences (patients) and the length of the sequence in nucleotides. Subsequent lines of the file contain the name of the sequence (patient ID in this case) followed by three spaces and the sequence.

```
   4 60
sequence 1    AAGCTTCACCGGCGCAGTCATTCTCATAAT
CGCCCACGGACTTACATCCTCATTACTATT
sequence 2    AAGNNNCACCGGCGCAATTATCCTCATAAT
CGCCCACGGACTTACATCCTCATTATTATT
sequence 3    AAGCTTCACCGGCGCAGTTGTTCTTATAAT
TGCCCACGGACTTACATCATCATTATTATT
sequence 4    AAGCTTCACCGGCGCAACCACCCTCATGAT
CGCCCACGGACTTACNNNNNNNNNNACTATT
```

FIGURE 2.10: **PHYLIP data file format.** An example sequence data file containing four sequences (each representing a different patient) each of 60 nucleotides in length over a whole single gene. Missing data is represented by Ns.

The control file (example in Appendix B) specified the names of the sequence data file, the tree structure file, and models and options for the analysis. In the control file, the variable *seqtype* was set to "1" (codonml), which carried out maximum likelihood analysis on codons of protein-coding DNA sequences using codon substitution models to detect positive selection. The non-default variables *model* and *NSsites* were set to "0" and "0 1 2" respectively, which implements a site model in which the ω ratio is allowed to vary among sites (among codons in this case) [Nielsen and Yang, 1998, Yang et al., 2000]. By entering three values for NSsites (0,1,2), three models are fitted to the same data: M0 (one ratio), M1a (nearly neutral) and M2a (positive selection) as is explained in Table 2.4. M1a and M2a are useful as a pair of models in forming a likelihood ratio test of positive selection. The more complex nested model M7 vs M8 is also used for the same comparison, which is a more powerful test than the M1a-M2a comparison which may be more accurate. However, the simpler model M1a-M2a was chosen in this analysis due to its faster run time and more robust framework. The control variable *codonFreq* was set to "2", which specifies the equilibrium codon frequencies in the codon substitution model, calculated from the average nucleotide frequencies at the three codon positions.

For each gene, PAML produced an omega (ω) ratio as a measure of the magnitude of selection and log likelihood values as a measure of the fit of the model. The ω ratio

TABLE 2.4: **Parameters in the site models used by codeml.** This table shows the parameters used for each of the three codeml models that have been specified in this PAML analysis. #p is the number of free parameters in the omega (ω) distribution. Parameters in parentheses are not free, so are not counted. In the likelihood ratio test comparing M1a (2 free parameters) against M2a (4 free parameters), the degrees of freedom (df) = 4-2 = 2.

Model	NSsites	#p	Parameters	References
M0 (one ratio)	0	1	ω	Goldman and Yang [1994], Yang and Nielsen [1998]
M1a (neutral + constraint)	1	2	p_0 (p_1 = 1 - p_0), $\omega_0 < 1$, $\omega_1 = 1$	Nielsen and Yang [1998], Yang et al. [2005]
M2a (M1a + positive selection)	2	4	p_0, p_1 (p_2 = 1 - p_0 - p_1), $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 1$	Nielsen and Yang [1998], Yang et al. [2005]

is calculated using evolutionary non-synonymous and synonymous substitution rates in codon model M0 (Equation 2.1), where K_a is the non-synonymous substitution rate (number of non-synonymous changes at non-synonymous sites) and K_s is the synonymous substitution rate (number of synonymous changes at synonymous sites). K_s acts as a proxy for neutral selection, as synonymous mutations are assumed to be selectively neutral. The maximum likelihood framework estimated two log likelihood (lnL) values, *lnL1* from model M1a and *lnL2* from model M2a. Model M1a models purifying selection as well as neutrality, constraining omega to less than or equal to 1 (Equation 2.2), whereas model M2a allows for positive selection as an additional parameter (Equation 2.3). Model M2a is parametised with the same values as model M1a but with added degrees of freedom to soak up residual variation.

$$\omega = \frac{K_a}{K_s} \tag{2.1}$$

$$\text{Model 1: } \omega <= 1 \tag{2.2}$$

$$\text{Model 2: } \omega <= 1 \mid \omega > 1 \tag{2.3}$$

During this analysis a problem occurred in which it was discovered that Model M2a was not always able to find the global maximum log-likelihood, sometimes reporting a local log-likelihood optima instead. This resulted in a negative difference between the log likelihoods of the two models, with the lnL value for M2a smaller than that of M1a. A negative delta-lnL is not possible if true optima have been found, since of the two nested models the one with two more parameters (M2a) should fit the data better than M1a resulting in a positive delta-lnL. A negative delta therefore means that a local optima has been found in the more complex model.

To overcome this problem and avoid obtaining a local log-likelihood optimum, the "codeml" analysis was adapted to run ten times over each gene, instead of being restricted to just one iteration. The least negative log-likelihood value for M2a from these ten iterations was assumed to be the global optimum and so was then used to calculate the p-value for that gene. However, the global log-likelihood was not always found after ten iterations. dchisq in R cannot cope with a negative delta so in these cases the negative differences were changed to 0 before a p-value was calculated, resulting in a very insignificant p-value which would not affect the significant results.

Codeml was adapted for use on cancer data. These innovations included using pairwise deletion at regions where data had been annotated as missing, by setting the *cleandata* option to "0", and including stop codon variants (stop-lost and stop-gained) in the model by recoding them as non-synonymous variants in the evolutionary analysis. The benefit of using pairwise deletion is that only sites which had missing characters in the pair were removed, rather than removing all sites involving ambiguity characters from all sequences using PAML's default of complete deletion (*cleandata* = "1"). This leaves more sequence information for PAML to work with which increases the power. Using this option was made possible by the fact that all sequences in this analysis were artificially generated and therefore treated as equally related, since cancer mutations

for each patient had been edited onto the same reference transcript for each gene, as opposed to using species data in which using pairwise deletion could lead to systematic biases. The ambiguity character that has been used in this analysis is "N", as is shown in Figure 2.10.

2.4.1.2 Limitations of PAML

Saturation and recombination can both be limiting factors when using codeml in PAML to estimate omega values. However neither are considered a problem in my dataset, since all alignments input into PAML are artificially generated by editing the cancer-specific SNVs onto a reference transcript for each patients, and hence all sequences are sufficiently related to the reference transcript to be aligned. As a result, there is a lack of divergence which can cause the saturation problem for which omega ratio estimations are commonly criticised. Additionally, no recombination is present in this dataset since cancer evolution is an example of clonal selection; this is why it was possible to edit the cancer-specific variants for each gene onto the same reference transcript for each patient.

2.4.2 Computational and statistical tests

2.4.2.1 P-value and FDR

As model M2a is nested within model M1a in codeml, we can calculate if the fit of the observed data to the more complex than M1a model (M2a) is better than is expected by chance. This is calculated as two times the difference in log-likelihood between models compared to a χ^2 distribution with two degrees of freedom (the difference in number of free parameters between models)[Yang et al., 2005]. This was implemented using the dchisq function in R version 3.0.0 [R Core Team, 2013], to generate a p-value for each gene as a measure of the significance of positive selection.

The set of p-values was then adjusted for multiple comparisons using p.adjust in R version 3.0.0 [R Core Team, 2013]. The particular method we used was the "BH" (aka "fdr") method of Benjamini and Hochberg [Benjamini and Hochberg, 1995], which controls the false discovery rate (FDR), the proportion of false discoveries expected amongst the rejected hypotheses. This produced a set of FDR values, one per gene, which was used as the final measure of significance for each gene accounting for multiple hypothesis testing. The false-discovery rate is a less stringent condition than the family-wise error rate (controlled by other correction methods such as the Bonferroni correction), so is more powerful as an adjustment method.

2.4.2.2 Log transformation of FDR

FDR values were plotted in R version 3.0.0 [R Core Team, 2013] against the omega estimates calculated in PAML for each gene, with FDR along the y-axis and omega along the x-axis. However, FDR values were log transformed for easier visualisation of significant genes with small FDR values using log10(-log10(FDR)), so that the most significant FDR values (smallest FDR values) were plotted at the top of the graph. The "double logging" was employed to stretch the less-significant end of the distribution which remains informative but is otherwise difficult to visualise. Logging presented a problem for any FDR values of zero, since log10(-log10(0)) equals infinity and so cannot be plotted. To overcome this, all FDR values of zero were changed to 1e-200 in R (version 3.0.0) [R Core Team, 2013], which is then small enough to show on the plot that these genes are highly significant.

2.4.2.3 Graphics

R (version 3.0.0) [R Core Team, 2013] package 'RColorBrewer' (version 1.0-5)[7] was used to highlight points in colour on the omega scatter plots in Chapter 4, Chapter 5 and Chapter 6 corresponding to significant genes after they had been plotted in R.

[7] http://colorbrewer.org

2.4.3 Preparation for functional sub-region analysis

As an initial test for functional sub-region analysis of short linear motifs in PAML, the TP53 gene was used as a prototype and was split into phosphorylation sites and non-phosphorylation sites using Perl (version 5.14.1)[8]. The perl script split the set of gene sequence alignments based on modification site annotations and produced two output files, one containing the concatenated sites of interest (phosphorylation sites), and the other containing the remainder of the gene. Gene codons were annotated using experimentally validated phosphorylation site coordinates taken from PhosPhoSite version 1.0 [Hornbeck et al., 2012]. PhosPhoSite version 1.0 [Hornbeck et al., 2012] is based on UniProt [UniProt Consortium, 2014] protein sequences, which was incompatible with our transcripts that were sourced from Ensembl. A coordinate transformation system was employed to overcome this annotation problem that used Exonerate (version 2.2) [Slater and Birney, 2005] to map the protein sequence from PhosPhoSite to the Ensembl sequence for the longest transcript we used. This was to ensure that the sequences matched, and in cases where they did not match a coordinate transformation system was implemented to transform the PhosPhoSite coordinates in order to realign the proteins to the Ensembl transcripts so that the correct genomic position for each site was reported.

Globular protein domain coordinates for all genes were obtained from Ensembl version 75 [Flicek et al., 2014] using Perl, in preparation for kinase domain analysis. These coordinates were then applied as vector annotations to the gene-based data of the protein kinase MAPK1 to split the data into kinase domain and remaining part of gene, in the same way as was done for the phosphorylation site data.

For both linear motif and globular domain prototype analysis, each file containing either the functional regions of the gene or the non-functional concatenated regions was then ready to be run through PAML for evolutionary analysis, to obtain two separate omega estimates for each gene.

[8] http://www.perl.org/

Chapter 3

Somatic single nucleotide variant analysis

3.1 Introduction

The calling of cancer-specific variants is challenging due to heterogeneity between samples and variability between analysis pipelines [Alioto et al., 2014]. In this chapter I characterise genome-wide SSNV patterns across different tissues of origin and compare results between my own calling of TCGA samples (n=1,005) with the variant calls of the Lawrence dataset (n=4,728). 701 patients with coding mutations overlap between each dataset (an additional patient is present in both datasets but does not harbour any coding mutations in the Lawrence dataset).

TABLE 3.1: **TCGA SSNV counts by cancer type.** For the heterozygous filtered TCGA cancer-specific SNVs, this table shows the total number of SSNVs over all patients for each cancer type and in the whole dataset of 1005 patients, as well as the mean number of SSNVs per patient for each cancer type and the whole dataset calculated by dividing the total number by the number of patients in that cancer type or dataset (rounded to the nearest whole number).

Cancer type	Total no. of SSNVs	Mean no. of SSNVs per patient
BLCA	4,458	297
BRCA	10,194	93
CESC	2,993	214
COAD	3,213	321
GBM	115,683	556
HNSC	12,372	146
KIRC	10,482	60
KIRP	894	56
LAML	760	14
LGG	12,789	256
LUAD	6,507	250
LUSC	16,665	314
OV	16,516	220
PRAD	1,080	28
STAD	4,164	219
THCA	207	11
UCEC	25,657	675
Whole dataset	**244,634**	**243**

3.2 Results

3.2.1 Summary variant statistics on a per patient basis

3.2.1.1 TCGA SNVs

Following filtering against control-specific and germline variants the TCGA dataset contains 244,634 heterozygous somatic (cancer-specific) single nulceotide variations (SSNVs); averaging at 243 SSNVs per patient (Table 3.1, Table 3.2). A breakdown of these SSNVs by cancer type is shown in Table 3.1. These mutations include 185,378 non-synonymous (missense) mutations and 59,256 synonymous mutations (Table 3.2).

TABLE 3.2: **TCGA SNV counts by mutation type.** For the heterozygous filtered TCGA cancer-specific SNVs, this table shows the total number of each mutation type in the whole dataset of 1,005 patients, as well as the mean number of each mutation type per patient calculated by dividing the total count for each class by 1,005 (rounded to nearest whole number). Stop-gained, stop-lost, start-lost and start-gained (non-synonymous start) have been included in the non-synonymous coding category. Synonymous-start and synonymous-stop have been included in the synonymous coding category.

Mutation type	Total no. of SSNVs	Mean no. of SSNVs per patient
Non-synonymous coding	185,378	184
Synonymous coding	59,256	59
Total	**244,634**	

Non-synonymous to synonymous mutations were present at a rate of 3.13 (2dp), where stop and start codon mutations were included as non-synonymous (stop and start codon mutations are included in later evolutionary analysis in calculation of omega ratios). Rates were calculated using values from Table 3.2.

3.2.1.2 Lawrence SNVs

Lawrence data was processed through the Broad Institute's filtering and annotation pipeline using their analysis platform "Firehose" to obtain a set of cancer-specific mutation calls. I filtered this dataset specifically for single nucleotide variants. The filtered Lawrence et al. [2014] dataset consists of 1,663,133 SSNVs, averaging at 352 SSNVs per patient (Table 3.3, Table 3.4). A breakdown of SSNVs by cancer type is shown in Table 3.3. These mutations include 520,588 non-synonymous (missense), 200,694 synonymous, 44,501 nonsense (stop-gained), 174 nonstop (stop-lost) and 897,176 non-coding mutations over all patients (Table 3.4). The 765,957 coding mutations in the Lawrence dataset have been highlighted in green in Table 3.4, and used for all subsequent analysis in this chapter (for direct comparison with coding TCGA SSNVs) and in the evolutionary analysis of subsequent chapters.

TABLE 3.3: **Lawrence SSNV counts by cancer type.** For the Lawrence cancer-specific SNVs, this table shows the total number of SSNVs over all patients for each cancer type and in the whole dataset of 4,728 patients, as well as the mean number of SSNVs per patient for each cancer type and the whole dataset calculated by dividing the total count by the number of patients in that cancer type or dataset (rounded to the nearest whole number). This dataset includes all coding and non-coding variants, however a subset of coding mutations only have also been shown. Cancer types highlighted in red are those that have lost patients with no coding SSNVs in the coding SSNVs dataset. 16 patients were lost altogether over six different cancer types.

Disease	**All SSNVs**			**Coding SSNVs only**		
	Patient number	Total SSNVs	Mean SSNVs per patient	Patient number	Total SSNVs	Mean SSNVs per patient
BLCA	99	30736	310	99	26673	269
BRCA	892	82889	93	888	41953	47
CARC	54	1604	30	54	1564	29
CLL	159	7002	44	159	2718	17
CRC	233	121831	523	233	77141	331
DLBCL	57	12574	221	57	10660	187
ESO	141	101303	718	141	17280	123
GBM	291	48654	167	291	19047	65
HNSC	384	67329	175	384	58756	153
KIRC	417	26335	63	417	22240	53
LAML	196	7450	38	194	3701	19
LUAD	404	235951	584	400	137207	343
LUSC	177	262148	1481	177	60314	341
MED	92	1261	14	92	1085	12
MEL	118	284456	2411	118	72226	612
MM	206	78910	383	205	10038	49
NB	76	1766	23	76	1515	20
OV	316	41543	131	316	17545	56
PRAD	137	29132	213	133	2513	19
RHAB	32	283	9	31	235	8
UCEC	247	219976	891	247	181546	735
Whole dataset	**4,728**	**1,663,133**	**352**	**4,712**	**765,957**	**163**

TABLE 3.4: **Lawrence SSNV counts by mutation type.** For the Lawrence cancer-specific SNVs, this table shows the total number of each mutation type in the whole dataset of 4,728 patients, as well as the mean number of each mutation type per patient calculated by dividing the total count for each class by 4,728 (rounded to nearest whole number). Coding SSNVs have been highlighted in green, and used for all subsequent analysis in this chapter. All other SSNVs in this table fall in non-coding regions that have been captured by targeted exome capture techniques prior to sequencing.

Mutation type	Total no. of SSNVs	Mean no. of SSNVs per patient
3' UTR	39,455	8
5' UTR	13,216	3
Intron	784,909	166
Non-synonymous	520,588	110
Non-coding transcript	1,789	0
Nonsense mutation	44,501	9
Nonstop mutation	174	0
Promoter	7,743	2
Splice site	31,130	7
Synonymous	200,694	42
Other	18,934	4
Total SSNVs	**1,663,133**	
Total coding SSNVs	**765,957**	

Using just the coding mutation counts (non-synonymous, nonsense, nonstop and synonymous), the ratio of non-synonymous (including nonsense and nonstop mutations) to synonymous mutations was estimated at a rate of 2.82 (2dp) (Table 3.4). This is lower than the non-synonymous:synonymous mutation ratio calculated for the TCGA dataset, indicating an increased number of non-synonymous SSNVs relative to synonymous SSNVs in the TCGA dataset compared to the Lawrence dataset.

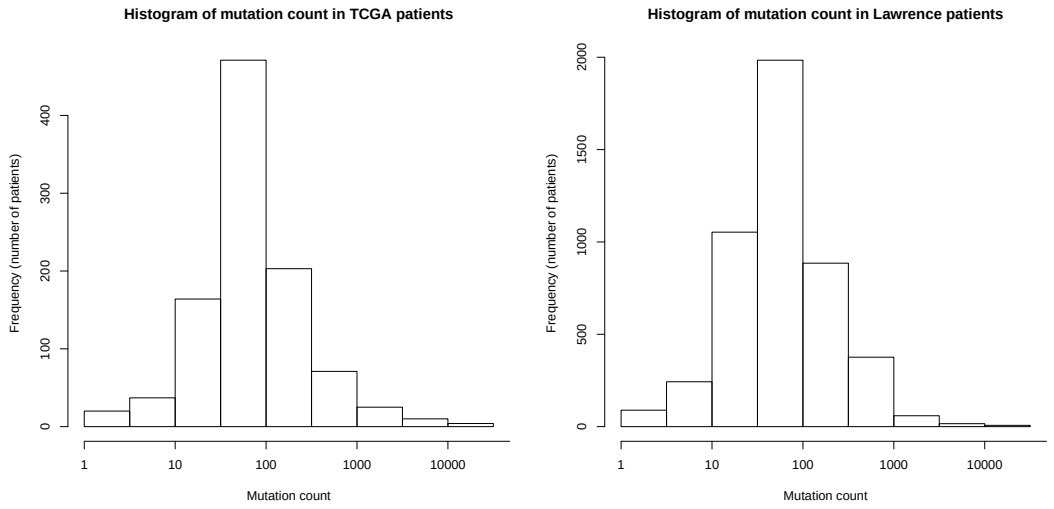
While the TCGA dataset consists of just heterozygous SSNVs (genotype of 0/1 in the tumour sample), filtering of the Lawrence cancer-specific mutations is not well documented. It is possible that they have not filtered out homozygous cancer-specific SNVs (with a genotype of 1/1 in the tumour sample).

In Table 3.4 nonsense and nonstop mutations are also known as stop-gained and stop-lost mutations, respectively. These stop codon mutations are often associated with a

loss of function and hence are commonly seen in tumour suppressor genes. Amongst the remaining coding base substitutions in Table 3.4, synonymous mutations do not alter the encoded protein so are not considered to have a functional effect on the encoded protein, and non-synonymous (missense) mutations change the encoded protein which can result in either activation (in oncogenes) or inactivation (in tumour suppressor genes) of the encoded protein.

In Table 3.4, 'non-coding transcript' refers to a mutation in a region of the genome that is transcribed into RNA but is not translated into a protein. In this table synonymous and silent mutations have been grouped together, since Lawrence et al. [2014] have classed some synonymous mutations as silent. However these terms are not interchangeable. Silent is the broad term for all DNA mutations without phenotype or consequence that do not change the protein and they can occur in either non-coding regions of the genome or in exons. Synonymous mutations on the other hand only occur in coding exons and do not change the encoded amino acid. However synonymous mutations are not always silent. For example a synonymous non-silent mutation could occur that does not alter the encoded amino acid but does knock out an exonic splicing enhancer.

It should be noted that the total number of SSNVs recorded in the Lawrence paper was 3,078,483, which is almost double the number of SSNVs (1,663,133) recorded in the Lawrence dataset used in this analysis (Table 3.3, Table 3.4). A possible explanation for this discrepancy could be that these missing variants were not in genes and were therefore not downloaded for this analysis, since the MAF files were downloaded on a per gene basis. However since Lawrence et al. [2014] used exome sequences, not many mutations would have been captured outside of the genic regions so these mutations would not account for much of the missing variation.



(A) TCGA (B) Lawrence

FIGURE 3.1: **Histograms of mutation count distribution over all patients in TCGA and Lawrence datasets.** Histograms showing the distribution of coding cancer-specific mutation counts for each patient over (A) whole TCGA dataset of 1,005 patients and (B) whole coding SSNVs Lawrence dataset of 4,712 patients. The x-axis denotes the SNV count on a log scale (log transformed for easier visualisation), but x-axis labels remain the original mutation count values. The y-axis represents the frequency of patients for each mutation count class. The area under the graph represents the total number of patients.

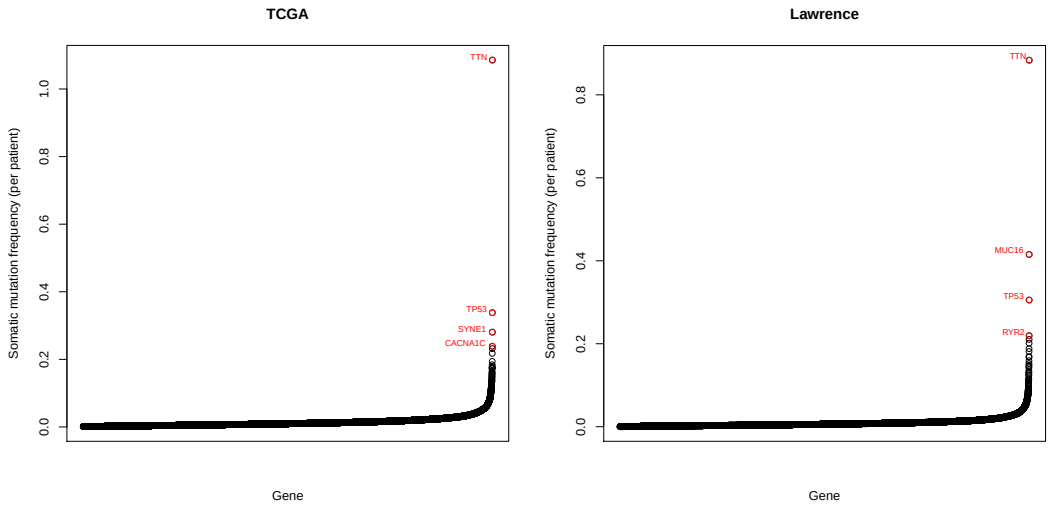
3.2.1.3 Comparison of TCGA and Lawrence datasets

In Figure 3.1, the coding mutation counts per patient over all tumour types have been compared for each dataset. As can be seen from the histograms, the distributions are very similar for both datasets, both displaying an approximately normal distribution with most patients having an intermediate mutation frequency. However, both distributions are slightly skewed to the left with a tail of very high mutation frequencies. This is most likely to be caused by the presence of UCEC patients in both the TCGA and Lawrence datasets that on average exhibit a higher SSNV rate than other tumour types (Table 3.3).

3.2.2 SSNVs on a per gene basis

Figure 3.2 shows the mutation frequency per patient for each gene in both the TCGA and Lawrence datasets. In both plots the relative coding SNP abundance when number of patients is taken into account has been shown, by dividing the total SNV count within each gene by the number of patients in the dataset. So in the TCGA plot, each gene's SNV count has been divided by 1,005 and in the Lawrence plot each SNV count per gene has been divided by 4,712. This allows a more direct comparison between datasets. Both plots show that TTN exhibits the highest mutation frequency of all genes in both the TCGA and Lawrence datasets. However TTN is known to be the largest gene in the human genome with an open reading frame of 107,976 nucleotides. The length of this gene could therefore introduce a bias, since length of gene has not been accounted for in these plots. Within this analysis all cancer types have been combined so that the TCGA and Lawrence plots consist of 17 and 21 different tumour types respectively. It should also be noted that although there is considerable overlap between TCGA and Lawrence patients with 701 coding patients over 11 tumour types overlapping both datasets, they are not identical sample sets, with varying cancer types and patient numbers per specific cancer type. These differences may influence the results seen in Figure 3.2, potentially generating a bias towards genes that are more frequently mutated in certain tumour types consisting of a larger proportion of patients. Consequently any differences cannot solely be attributed to differences in pipelines.

In Figure 3.3, gene length has also been accounted for by dividing the average SSNV count per patient for each gene by the length of the gene in nucleotides (using the longest CDS length for each gene). By running this gene length correction, TTN is removed from the top most frequently mutated genes. TP53 is now shown to contain the most mutations (per nucleotide) of all genes in both the TCGA and Lawrence datasets. TP53 is known to be the most mutated gene in cancer, supporting such results. PTEN and VHL are also detected as being highly significantly mutated in both the TCGA and Lawrence datasets, both of which are known cancer genes implicated in tumour progression, with PTEN known to be causally implicated in glioma, prostate

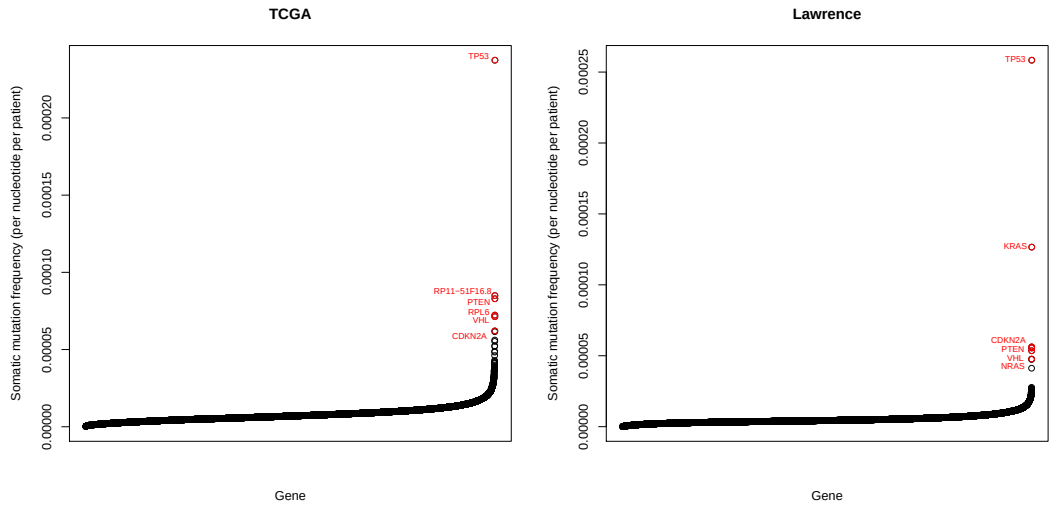


(A) TCGA (B) Lawrence

FIGURE 3.2: **Mutation frequency on a per gene basis.** For each gene, the coding mutation frequency averaged over all patients has been plotted along the y-axis for (A) the whole TCGA dataset and (B) the whole coding SSNVs Lawrence dataset of 4,712 patients. Genes with the highest mutation frequencies have been highlighted in red.

and endometrial cancers and VHL known to be a driver gene in renal cancers [Futreal et al., 2004].

A problem encountered in this gene-based analysis was that some genes did not have a CDS length recorded in Biomart. It is possible to find non-coding transcripts for a gene (nonsense-mediated mRNA decay (NMD) targets) annotated in Ensembl, which could be the source of genic transcripts without CDS measures. However, since there are usually multiple transcripts for a gene, it would be expected that at least one transcript contained a CDS for each protein-coding gene. Therefore this was an unusual finding. Furthermore, some gene IDs were not reported in Biomart at all (with or without a CDS length).



(A) TCGA (B) Lawrence

FIGURE 3.3: **Mutation frequency on a per gene basis with gene length normalisation.** For each gene, the coding mutation frequency per nucleotide averaged over all patients has been plotted along the y-axis for (A) the whole TCGA dataset and (B) the whole coding SSNVs Lawrence dataset of 4,712 patients. The length of the gene has been accounted for by dividing the mutation frequency per patient for each gene by the length (in nucleotides) of the longest CDS (portion of the mRNA transcript that is translated into protein) of that gene. Genes with the highest mutation frequencies per nucleotide have been highlighted in red.

3.2.3 Patients overlapping datasets

3.2.3.1 Mutation frequency comparison

701 patients with coding SSNVs are present in both the TCGA and Lawrence datasets. For each of these patients I plotted the mutation frequency per Mb in the TCGA dataset against the mutation frequency per Mb in the Lawrence dataset in Figure 3.4. I observed a strong positive linear correlation between the TCGA and the Lawrence mutation frequencies (Pearson's cor=0.82, p-value <2.2e-16), illustrating a strong concordance between the shared patients in the two datasets, with each analysis pipeline identifying a similar total number of SSNVs per patient.

However, there seems to be a small subset of patients that are right-shifted in Figure 3.4, exhibiting a markedly higher mutation frequency in TCGA compared to that

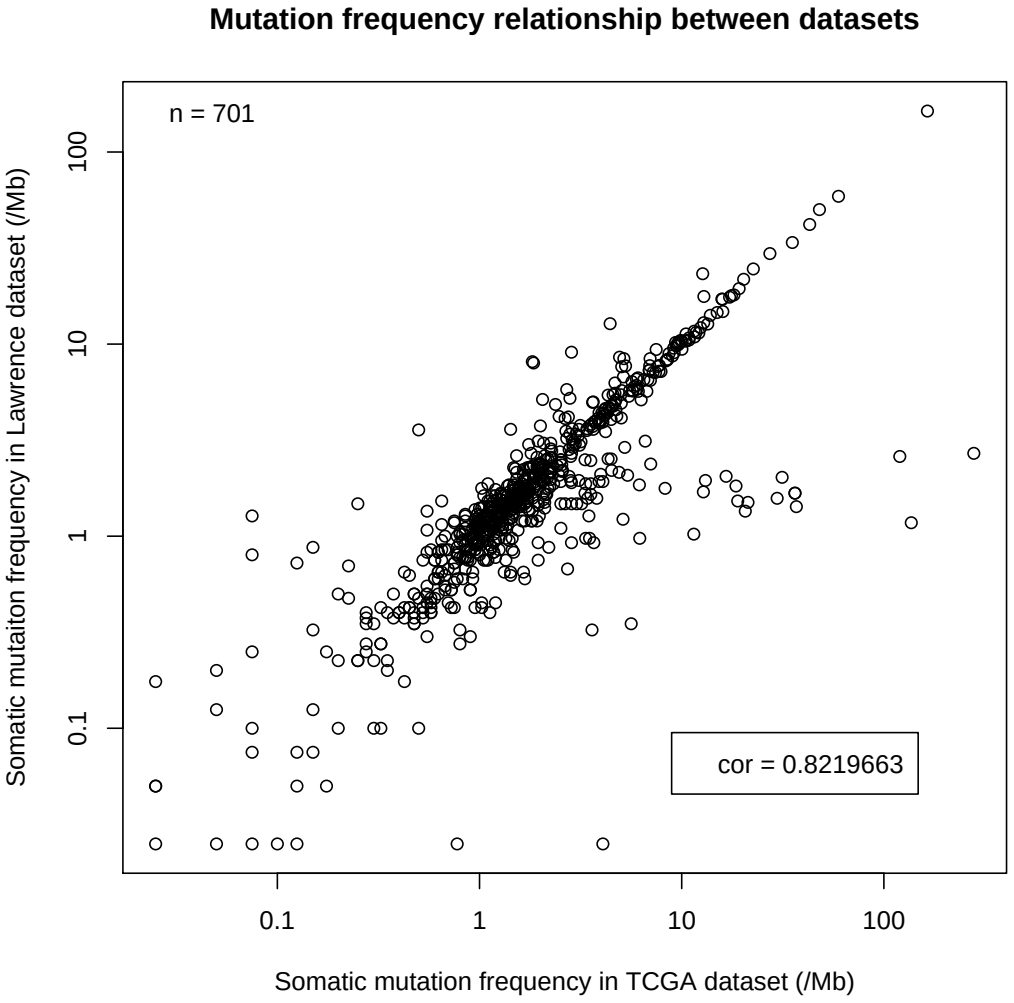


FIGURE 3.4: **Mutation frequency relationship between datasets for shared patients.** The TCGA and Lawrence mutation frequencies per Mb have been plotted for the 701 patients that are present in both the TCGA and Lawrence datasets, to investigate the concordance between the two different pipelines. Both the x-axis and y-axis has been log transformed for clearer visualisation of the correlation, however the axis labels are the true mutation frequency values. For each patient, the mutation frequency in the TCGA dataset is plotted along the x-axis and the corresponding mutation frequency in the Lawrence dataset is plotted along the y-axis.

of Lawrence. These patients were identified as outliers in Figure 3.5. Colour-coding of these 24 patients reveal that multiple cancer types are encompassed by these outlier patients (GBM, COAD, BRCA, OV and KIRC) however most (n=16; highlighted in red) are of the tumour type GBM (total number of GBM patients shared by both datasets = 171). It was important to investigate this discrepancy as it could indicate that either a systematic lack of sensitivity in the Lawrence calling pipeline or systematic over-calling in my TCGA pipeline was responsible for skewing results, either of which would have a detrimental effect on downstream analysis.

To investigate the group of 24 outliers further, the mutation spectrum of this subset compared to the rest of the set was examined in Figure 3.6 and Figure 3.7, to decipher if these 24 patients have a specific mutation spectrum in the TCGA dataset compared to the Lawrence dataset, which could help elucidate the source of the observed high mutation frequency in TCGA in terms of how SNVs have been called differently between the two pipelines.

The group of 701 overlapping patients was split into two subsets: the 24 outliers and the 677 non-outliers (remaining patients). The proportions of the six different possible base-pair substitutions for each of the 24 outlier patients was plotted in Figure 3.6, separately for the variants called in the TCGA dataset and the variants called in the Lawrence dataset. The same was done for the non-outlier patients in Figure 3.7.

The mutation spectra for the TCGA outlier subset in Figure 3.6 is more uniform than the mutation spectra observed in Lawrence for the outlier subset. The Lawrence outlier subset is also more similar to both the TCGA and Lawrence non-outlier subsets in Figure 3.7, with much higher C→T substitutions than any other mutation class, although the spectra in Figure 3.7 show much more variation in base-pair substitutions across all patients, but this is expected since more patients were used for these plots and because many patients from multiple cancer types have been grouped together that are assumed not to have a specific mutation spectrum. This result leads to the speculation that the TCGA variant calling pipeline is miss-calling SNVs, and calling more SNVs generally of all types of mutation making the spectrum more uniform across all

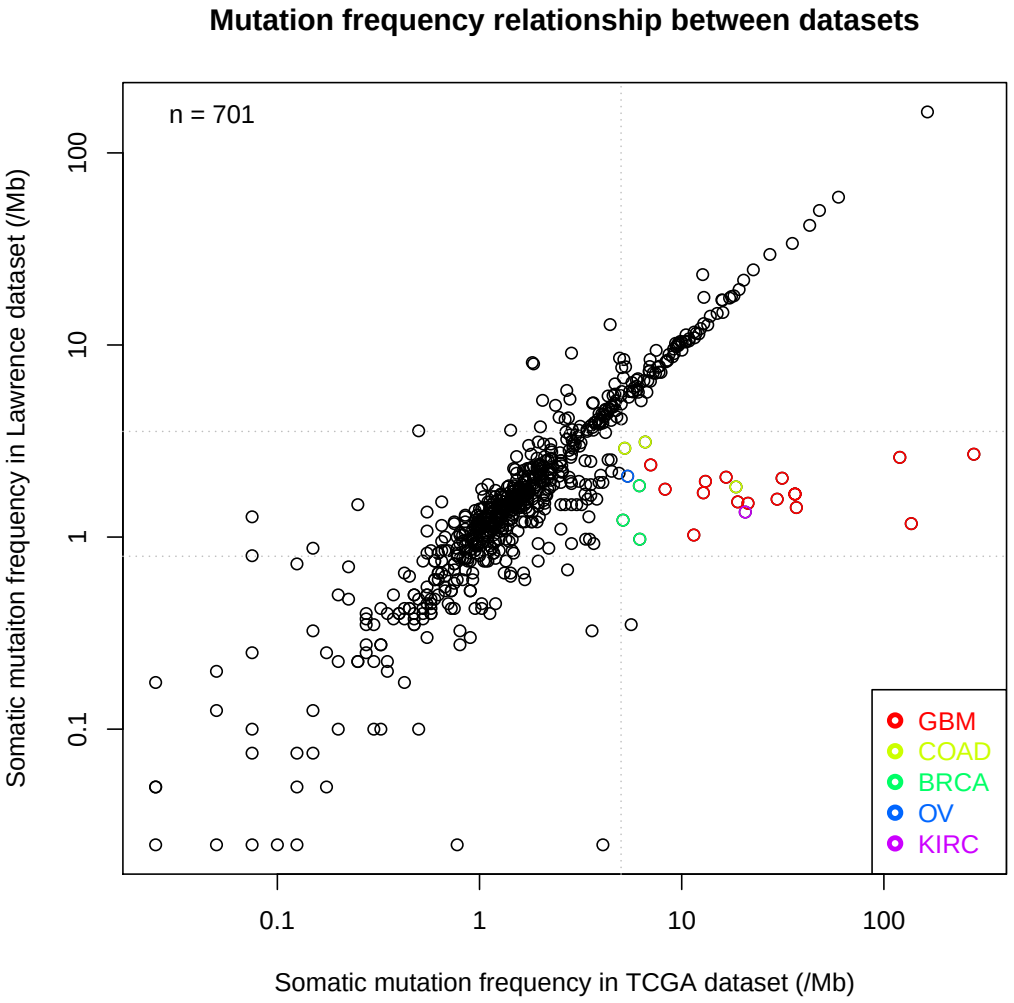
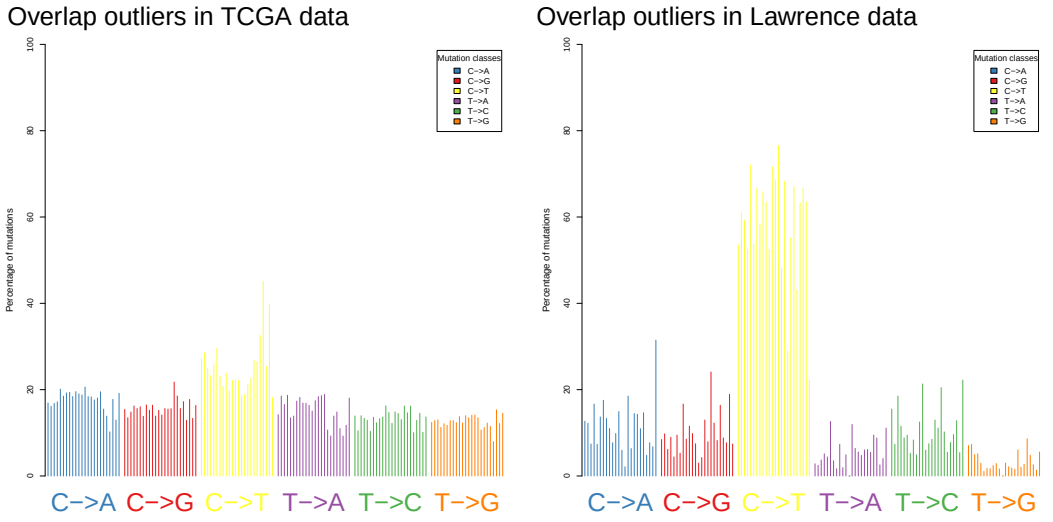


FIGURE 3.5: **Mutation frequency relationship outliers.** A subset of 24 outliers each with a higher than expected TCGA mutation frequency compared to the mutation frequency in the Lawrence dataset observed in Figure 3.4 have been highlighted here. Dotted grey lines have been drawn on the plot to show the range of the mutation frequencies of these patients. The patients have been colour-coded according to their cancer type, shown in the legend.

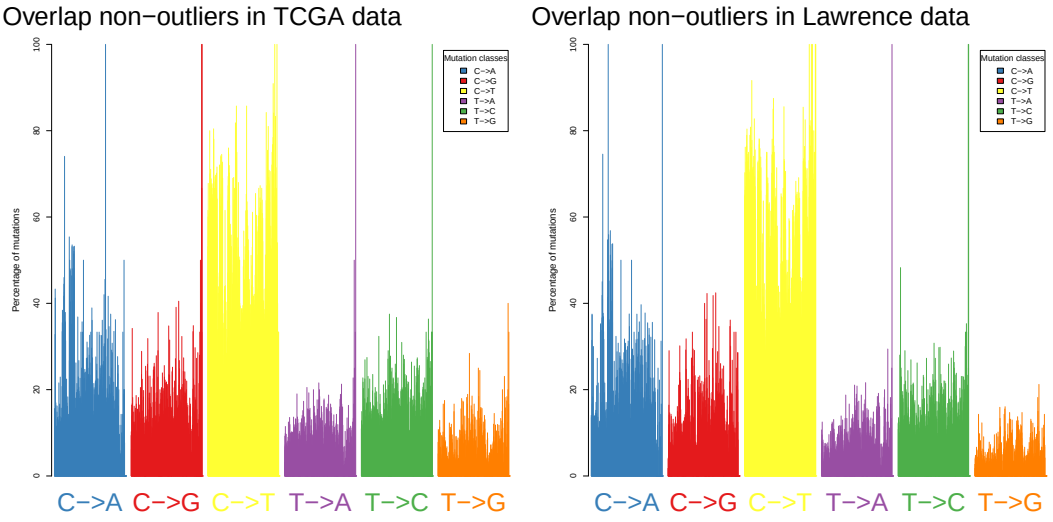


(A) TCGA (B) Lawrence

FIGURE 3.6: **Mutation spectrum of patients in outlier subset from patients overlapping both datasets.** Outlier subset consists of 24 patients. For each patient the proportion of each mutation class has been plotted along the y-axis in (A) the TCGA dataset and (B) the Lawrence dataset.

six base-substitution classes. I speculated that this may be due to increased INDELs in these patients, which can lead to spurious SNV calls around INDELs due to the misalignment of flanking sequence. The TCGA pipeline did not specifically exclude SSNV calls proximal to INDELs. It is undocumented whether the Lawrence pipeline did.

To investigate this in just the 16 GBM patients, the mySQL TCGA database (tcga_pair_exome) of variants was mined to look for a possible elevated rate of INDEL calls in these 16 GBM outlier patients. Within the subset of 16 GBM patients, the mean number of INDELs per patient was found to be 16,289 (rounded to the nearest whole number), compared to 202 INDELs per patient in the non-outlier subset containing 685 patients (Figure 3.8). These two subsets were also compared to the average INDEL rate over the whole TCGA dataset of 1,005 patients (1,037 INDELs per patient), and also over the whole group of 701 overlap patients present in both TCGA and Lawrence datasets (569 INDELs per patient). Figure 3.8 clearly displays an inflated INDEL rate in the subset of 16 GBM outlier patients compared to the other 685 patients that overlap both



(A) TCGA (B) Lawrence

FIGURE 3.7: **Mutation spectrum of patients in non-outlier subset from patients overlapping both datasets.** Non-outlier group consists of all 677 patients not present in the outlier subset. For each patient the proportion of each mutation class has been plotted along the y-axis in (A) the TCGA dataset and (B) the Lawrence dataset.

datasets. Therefore, this higher rate of INDELs could explain the high rate of SNVs that are being called in these 16 GBM patients, as a result of over-calling SNVs next to INDELs. The INDEL rate in the 16 GBM patients was also compared to that of the other 192 GBM patients present in the whole TCGA dataset, to ensure that the high INDEL rate observed in the 16 GBM patients is specific to these 16 patients and is not common to all GBM patients (which would not explain the high rate of SNVs in just these 16 patients). Figure 3.8 shows that, as suspected, the INDEL rate is considerably inflated in the 16 GBM outlier patients compared to both the whole GBM subset of 208 patients (3374 INDELs per patient) and the subset of 192 GBM patients exclusive of the 16 outlier patients (2,298 INDELs per patient). Although it does seem that GBM patients exhibit a slightly increased frequency of INDELs compared to the rest of the TCGA patients when the 16 outlier patients are excluded, however a significant increase in INDELs is observed in the 16 outliers. This result confirms that a higher frequency of INDELs in the 16 GBM outlier patients is likely to be the cause of the high rate of called SNVs in these same patients.

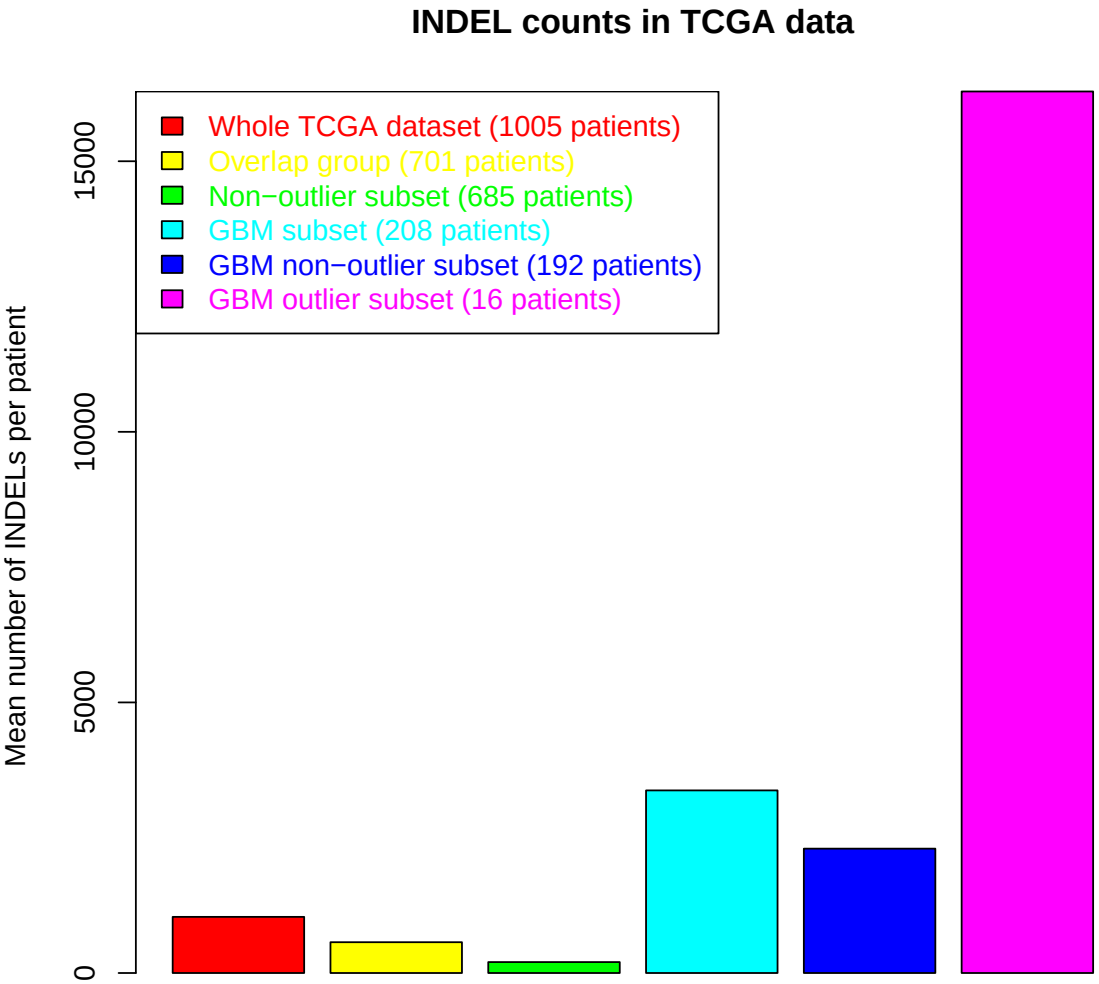


FIGURE 3.8: **INDEL counts in TCGA dataset.** This bar plot shows the mean number of cancer-specific heterozygous INDELs per patient in the 16 GBM outlier patients compared to the INDEL counts in: the whole TCGA dataset (1,005 patients), the overlap group of 701 patients, the 685 non-outlier patients in the overlap group (overlap patients excluding the 16 GBM outliers), all 208 GBM patients in the TCGA dataset and the 192 GBM patients that are not present in the GBM outlier subset of 16 patients. INDELs were only included in each of these four datasets if the cancer and control genotype quality >= 30. Includes only coding INDELs (since the "consequence" table was used to mine the *tcga_pair_exome* database, which contains only coding variants).

3.2.3.2 Concordance of variants detected

As a powerful way to show the differences and similarities between the two analyses, a QC check was performed to show how much variation was shared between the TCGA and Lawrence analyses for the 701 overlapping patients.

117,289 variants were detected over the 701 overlapping patients in the TCGA pipeline in this project, and 114,753 variants were detected in these same patients through the Lawrence pipeline. 71,845 of these variants concur in both the TCGA and Lawrence analyses, resulting in 61% (71,845 of 117,289) of the total TCGA variants being the same as 63% (71,845 of 114,753) of the total Lawrence variants. The average concordance of detecting the same variant in both analyses was therefore estimated as ∼62%.

Variants common to both analyses were found in 688 of the 701 patients, resulting in 13 patients (2%) that did not exhibit any shared variants between the two analyses.

3.2.4 Mutation spectra

The mutation profile can be defined as the mutational pattern incorporating information such as the numbers of each class of mutation, the DNA sequence flanking each mutated base and in transcribed regions whether the transcribed or untranscribed strand is preferentially mutated [Stratton, 2011]. Different mutation spectra may indicate differing underlying biology to tumours [Greenman et al., 2007, Pleasance et al., 2010a,b] and could influence susceptibility of different genes to subsequent mutational perturbation. There are potentially many alternate ways to define a mutation spectrum. For the purposes of this work I define the mutation spectrum based on the relative proportions of different nucleotide substitutions.

3.2.4.1 TCGA

Table 3.5 shows the relative proportions of the six different classes of nucleotide substitutions in the TCGA dataset, for each of the 17 tumour types. The mean proportions of

mutations have not been adjusted to account for the normal proportions of nucleotide changes across the genome. However, for the purposes of tumour classification based on mutation spectrum, as all tumours are assayed for mutational changes over a common interval (the targeted exome) this simple measure serves well for classification and comparison between patients. A higher C→T rate relative to all other mutation types is seen in the normal human genome and in cancer, and is also observed in Table 3.5 (highest proportions for each disease are highlighted in blue). However, for lung cancers (LUAD and LUSC) C→A/G→T transversions are observed as the commonest change. This is expected, as lung cancer is known to have a specific mutational signature defined by elevated C→A/G→T substitutions caused by tobacco exposure [Pleasance et al., 2010b].

In Table 3.6 the transition:transversion ratios have been calculated for each cancer type. In germline cells and in cancer, transitions are usually more common than transversions [Rubin and Green, 2009], which is the case for most cancers assayed here, particularly STAD which displays the highest number of transition mutations (A→G/T→C and C→T/G→A) relative to transversions. However, GBM, KIRC, LUAD, LUSC and OV all exhibit an increased rate of transversions (A→C/T→G, A→T/T→A, C→A/G→T and C→G/G→C) compared to transitions. For LUAD and LUSC this can be explained by the increased frequency of C→A/G→T caused by tobacco in lung cancers.

Figure 3.9 displays the distribution of mutation frequencies and spectra over the 17 different tumour types present in the TCGA dataset of 1,005 patients. LUAD has the highest median mutation frequency, which could be attributed to the mutator phenotype caused by exposure to tobacco in lung cancer.

TABLE 3.5: **Mutation profile proportions in TCGA dataset.** For each of the 17 cancer types in the TCGA dataset, the mean proportion ± standard deviation (3dp) per patient has been tabulated for each of the six mutation classes. The proportion has been calculated as a proportion of the total number of variants in that cancer type for that patient. For each patient the proportions were calculated and then the mean and standard deviation was taken over all patient proportions for each cancer type. The highest proportion for each cancer type is highlighted in blue.

Disease	Mutational signature					
	A⇒C/ T⇒G	A⇒G/ T⇒C	A⇒T/ T⇒A	C⇒A/ G⇒T	C⇒G/ G⇒C	C⇒T/ G⇒A
BLCA	0.044±0.021	0.115±0.078	0.050±0.035	0.118±0.051	0.218±0.129	0.455±0.124
BRCA	0.064±0.050	0.132±0.058	0.076±0.046	0.150±0.068	0.136±0.072	0.442±0.106
CESC	0.039±0.044	0.046±0.033	0.016±0.013	0.087±0.028	0.229±0.106	0.582±0.075
CRC	0.076±0.033	0.124±0.032	0.090±0.040	0.155±0.025	0.101±0.048	0.454±0.117
GBM	0.065±0.050	0.128±0.082	0.076±0.064	0.117±0.065	0.109±0.100	0.505±0.192
HNSC	0.033±0.023	0.106±0.050	0.059±0.040	0.180±0.131	0.144±0.077	0.478±0.148
KIRC	0.087±0.046	0.163±0.062	0.110±0.057	0.189±0.068	0.122±0.062	0.329±0.084
KIRP	0.102±0.052	0.191±0.069	0.110±0.068	0.144±0.074	0.115±0.052	0.338±0.159
LAML	0.058±0.093	0.165±0.161	0.062±0.147	0.108±0.116	0.120±0.164	0.487±0.227
LGG	0.056±0.069	0.177±0.108	0.038±0.039	0.090±0.070	0.081±0.064	0.557±0.149
LUAD	0.045±0.042	0.091±0.059	0.079±0.035	0.356±0.155	0.136±0.060	0.294±0.117
LUSC	0.032±0.012	0.109±0.044	0.082±0.029	0.348±0.115	0.140±0.057	0.289±0.102
OV	0.100±0.032	0.145±0.035	0.120±0.034	0.189±0.053	0.163±0.040	0.282±0.070
PRAD	0.070±0.051	0.131±0.161	0.056±0.083	0.151±0.122	0.090±0.063	0.503±0.126
STAD	0.073±0.048	0.106±0.050	0.060±0.040	0.127±0.043	0.075±0.033	0.559±0.089
THCA	0.087±0.098	0.153±0.111	0.162±0.150	0.090±0.104	0.079±0.082	0.430±0.183
UCEC	0.046±0.037	0.134±0.078	0.029±0.028	0.135±0.063	0.062±0.053	0.594±0.109

TABLE 3.6: **Transition:transversion ratios in TCGA data.** For each of the 17 cancer types in the TCGA dataset, the transition:transversion ratio (3dp) has been calculated by dividing the total number of transition mutations over all patients in a cancer type by the number of transversions for that cancer type. The number of patients in each subset has not been accounted for since that is not necessary when estimating ratios. The cancer type with the highest ratio is highlighted in red.

Cancer type	TS:TV
BLCA	1.788
BRCA	1.549
CESC	1.324
CRC	1.455
GBM	0.787
HNSC	1.167
KIRC	0.858
KIRP	1.032
LAML	1.912
LGG	2.979
LUAD	0.457
LUSC	0.559
OV	0.673
PRAD	1.628
STAD	3.342
THCA	1.379
UCEC	2.702

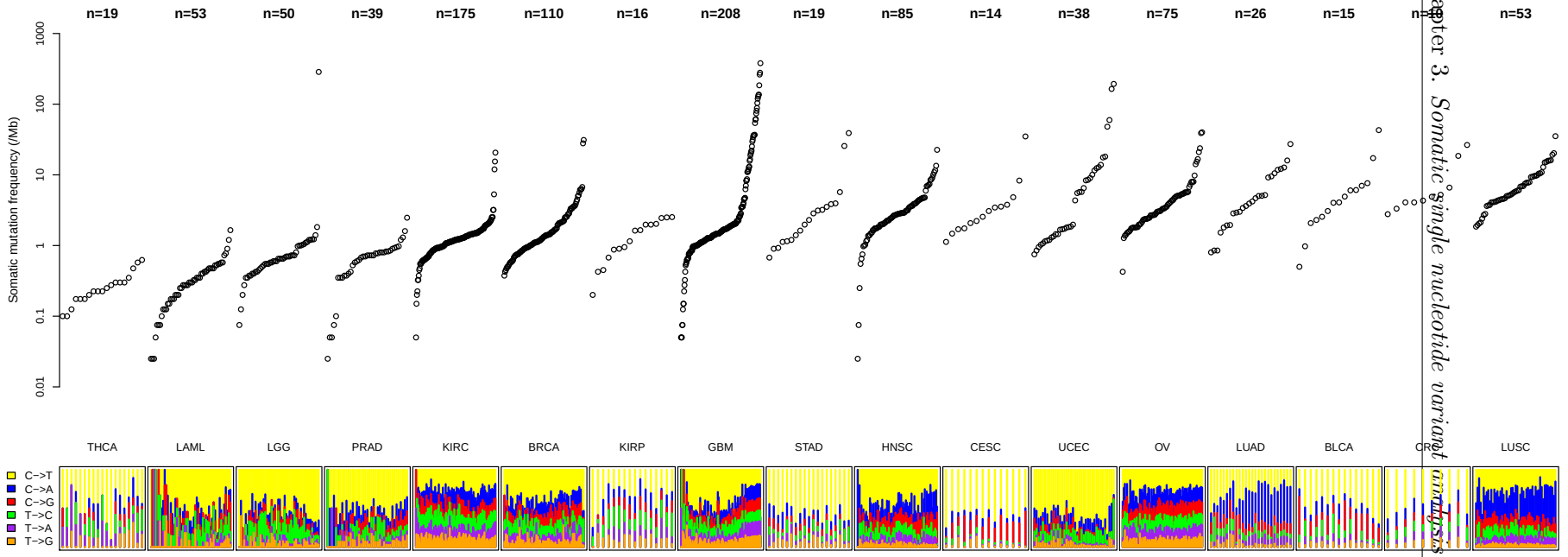


FIGURE 3.9: **Distribution of mutation rates and spectra across tumour types in the TCGA dataset.** Using the filtered set of cancer-specific heterozygous SNVs (non-synonymous and synonymous), the mutation frequency and spectra has been plotted for each of the 17 tumour types over 1,005 patients. Each dot represents a single patient, with the y-axis value indicating the total frequency of somatic SNVs in the exome per Mb. Tumour types have been ordered by their median somatic mutation frequency in ascending order. Within each tumour type, individual patients have also been ordered in ascending somatic mutation frequency along the x-axis. The mutation rates have been log transformed before plotting to spread out the data points for clearer visualisation, but the annotated y-axis values are the actual frequencies. The lower panel shows the relative proportions of the six different possible base-pair substitutions, as indicated in the legend. Sample sizes (patient numbers for each tumour type) are shown along the top of the plot. *R code used to generate plot can be found in Supplementary Appendix C where plot has been re-implemented from scratch.*

An interesting pattern is observed betweethe comparison of GBM patients in Figure 3.9 and in the corresponding distribution in the Lawrence dataset (Figure 3.14). The mutation spectra of GBM differs between the two datasets, and is an obvious outlier. In TCGA there is a sub-group of patients with very high mutation frequencies and a very different mutation profile compared to the rest of the GBM patients. This pattern is not observed in the Lawrence dataset. Since all GBM patients used in the TCGA analysis were available to the Lawrence study, it was speculated that the GBM patients with high mutation frequency and differing mutation profile in the TCGA dataset were for some unknown reason excluded from the Lawrence dataset, and therefore not present in the Lawrence dataset. To find out whether this was true, the patients common to both datasets were identified. Of the 208 TCGA GBM patients, 171 are common to both datasets, leaving 37 patients that are only present in the TCGA dataset. These 37 patients have been highlighted in black in Figure 3.10, showing that they are indeed predominately the patients with the highest mutation frequencies and unique mutation spectrum, explaining the difference between GBM in Figure 3.9 and Figure 3.14.

The question is then why were these 37 patients excluded from the Lawrence GBM subset, when they were available for download at the time of the Lawrence et al. [2014] study? Did they use different inclusion criteria? First, the tumour subtype was looked into. Of the four distinct subtypes of glioblastoma (GBM): proneural, neural, classical and mesenchymal, these 37 patients do not fall into a specific subtype and in fact cover all four subtypes, so this cannot be the exclusion criteria. Second, the sample type was investigated. 36 of the patients are primary tumours with blood derived normal matched samples, and one is primary tumour with solid tissue sample. And the other 171 patients overlapping both datasets also use both primary tumour, blood derived normal and solid tissue sample, so they cannot have filtered based on sample type. Another possibility was that exclusion was based on the tissue source center. However, again the 37 patients were from a variety of tissue source centers.

It is possible that these 37 GBM patients were excluded by Lawrence et al. [2014] due to their distinctly different mutation spectra and high mutation rates compared to the rest

of the GBM patients. Perhaps these patients have a high rate of INDELs and therefore were discarded due to the difficulty in calling SNVs near INDELs. The 16 GBM patients with high INDEL rates from the outlier set within the overlap patients also have a very high mutation frequency (brown dots in Figure 3.10). This also matched up with the mutation spectrum observed in the group of 16 GBM outliers (Figure 3.6) which is more uniform than the other patients, due to general over-calling. However if these 37 patients (black dots) were excluded based on high INDEL rates then it should follow that the 16 GBM patients would also be excluded from the Lawrence dataset, however they are present in both datasets. TCGA have filtered their tumour samples based on necrosis and heterogeneity data using stringent criteria (allowing samples with less than 50% necrosis and at least 80% tumour nuclei), so these 37 patients clearly have not been excluded due to high levels of necrosis.

The mutation spectra proportions have been displayed as a bubble plot in Figure 3.11, using all 1,005 patients in the TCGA dataset. Of the 12 possible mutation changes, the highest proportions are C→T and G→A transitions. This supports what is seen generally in the human genome in evolution, since C→T is the most common point mutation, corresponding to G→A occurring on the complementary strand of DNA, which is why the proportions are identical for both C→T and G→A. In cancer however, this pattern can be altered due to exposure to carcinogens such as tobacco smoke in lung cancer and UV radiation in melanoma, which is what has been investigated in this chapter by looking at the varying mutation spectra across both the TCGA and Lawrence datasets in specific tumour types.

From the TCGA data I used 53 LUSC patients and 38 UCEC patients to generate two further bubble plots (Figure 3.12), to inspect the mutation profile of tumour types more closely and observe differences between tumour types. LUSC is known to have a specific mutation spectrum caused by smoking, demonstrated by an elevated proportion of C→A (G→T on the complementary strand) transitions [Pleasance et al., 2010b], a pattern induced by tobacco's carcinogens. This is reflected in my results and provides evidence that the mutational spectrum of specific cancer types is influenced by

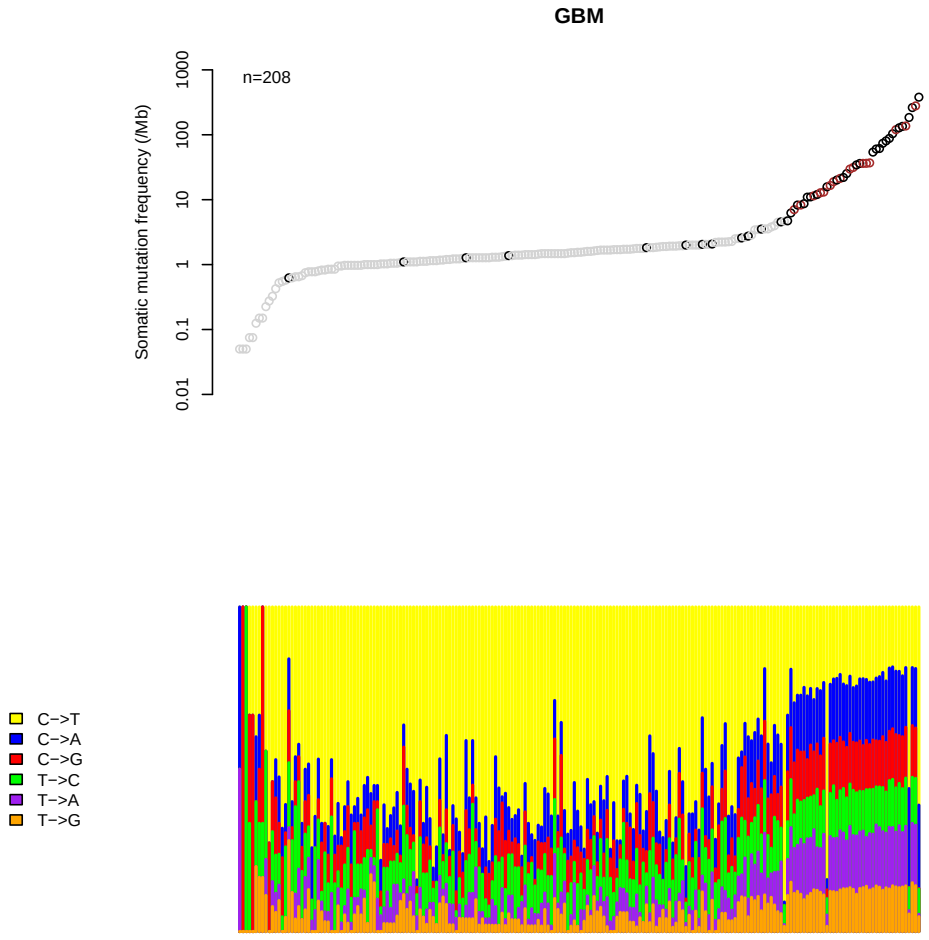


FIGURE 3.10: **Mutation rates and spectra for TCGA GBM patients excluded from Lawrence dataset.** This plot has been taken from Figure 3.9 to investigate the interesting mutation spectra observed in GBM in the TCGA dataset (compared to the GBM mutation spectra in the Lawrence dataset). Dots highlighted in black represent the 37 patients that are present in the TCGA dataset only. All other patients (grey dots) are present in both TCGA and Lawrence datasets. Brown dots represent the 16 GBM patients present in the overlap outlier set in Figure 3.5, so these patients are present in both the TCGA and Lawrence dataset.

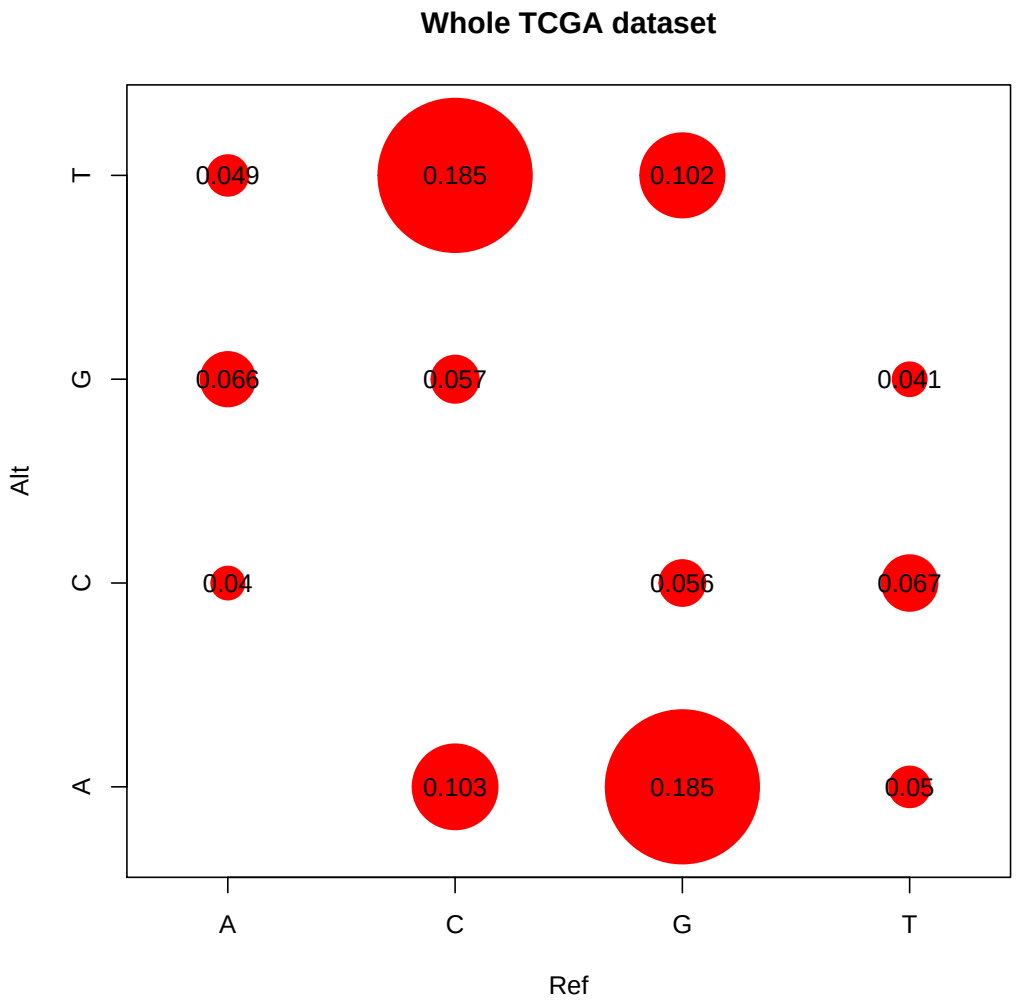
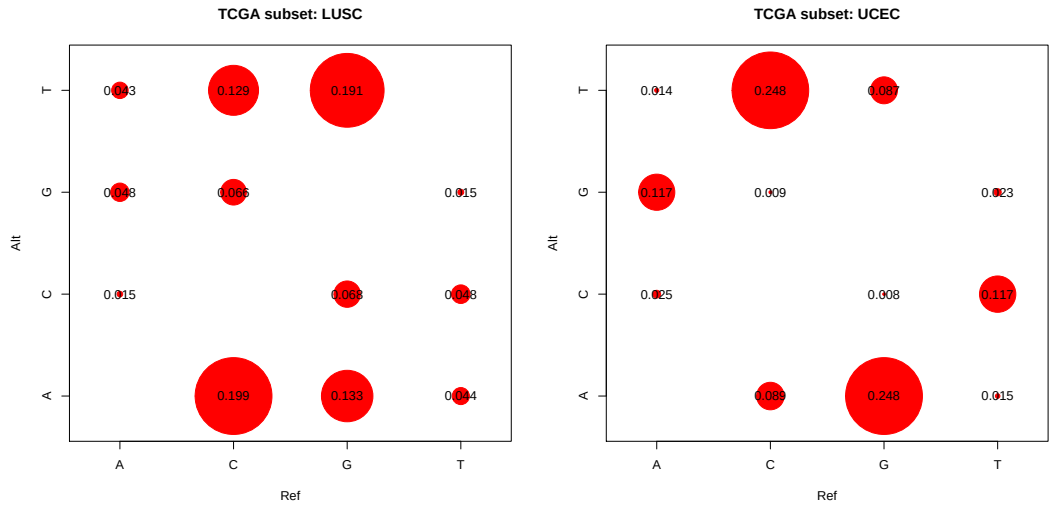


FIGURE 3.11: **Mutation spectra bubble plot for whole TCGA dataset.** For whole TCGA dataset over all 17 tumour types and 1,005 patients, the proportions of the 12 different possible base-pair substitutions have been shown, with the four possible reference bases along the x-axis and the four possible alternate bases (that the reference base has mutated to in the cancer) along the y-axis. The size of the red circles represent the proportion of nucleotide substitutions from the total number of cancer-specific mutations. Raw counts over whole dataset have been used to calculate proportions, rather than mean values per patient. For each mutation class, the proportion was calculated by dividing the mutation count for that mutation class by the total number of mutations in the dataset.



(A) LUSC (B) UCEC

FIGURE 3.12: **Mutation spectra bubble plots comparison by cancer type for LUSC and UCEC in TCGA dataset.** For (A) all 53 LUSC patients and (B) all 38 UCEC patients in the TCGA dataset, the proportions of the 12 different possible base-pair substitutions have been shown, with the reference allele along the x-axis and the alternate allele along the y-axis. The size of the red circles correspond to the size of the proportion. Raw counts over whole dataset have been used to calculate proportions, rather than mean values per patient. For each mutation class, the proportion was calculated by dividing the mutation count for that mutation class by the total number of mutations in the dataset.

environmental exposures. It can be seen that there is variation between the mutation profiles of LUSC and UCEC, with C→T (G→A) being the most common change in UCEC patients, however this is also the most common substitution observed in most cancer types [Kandoth et al., 2013].

As illustrated here, mutational signatures have been shown to vary depending on different environmental exposures. For example, Greenman et al. [2007] has shown substantial variation in the mutational profiles of 210 diverse cancers reflecting the different exposures acting on the different cancers. To investigate how mutation spectra vary across tumour types in the TCGA dataset, a hierarchical cluster tree and heatmap was generated in Figure 3.13 for the 1,005 patients in the TCGA dataset. Patients were clustered according to their single nucleotide mutation spectrum. Interestingly, the colour-coded panel corresponding to tumour type across the top of the heatmap

shows that these patient clusters are not specific to certain tumour types, as might be expected. So patients with the same tissue of origin do not necessarily have similar mutation spectra, suggesting there is minimal correlation between mutation spectra and tissue of origin.

3.2.4.2 Lawrence

Table 3.7 shows the relative proportions of the six different classes of nucleotide substitutions in the Lawrence dataset, for each of the 21 tumour types. As in the TCGA dataset, LUAD and LUSC are shown to exhibit more C→A/G→T mutations. However, interestingly NB also shares this mutation profile, suggesting that neuroblastoma and lung cancers could be caused by a similar mutational mechanism perhaps as a result of parental smoking.

MEL has been shown to harbour the highest number of transitions relative to transversions in Table 3.8, with a much higher transition rate compared to other tumour types, which is a known signature of UV exposure [Pleasance et al., 2010a].

Figure 3.14 displays the distribution of mutation frequencies and spectra over the 21 tumour types in the Lawrence dataset. In contrast to Figure 3.9, for the cases where the same tumour type is present in both datasets, Figure 3.14 and Figure 3.9 do not shown identical results. However, the numbers are variable for each of these cancer types between datasets and the TCGA data contains fewer patients than the Lawrence data, which could explain the differences. The biggest difference is observed in GBM, which has previously been investigated in Figure 3.10. Figure 3.14 replicates what has already been performed in the Lawrence et al. [2014] supplementary material. Although these two plots do not look identical despite using the same data, a direct comparison cannot be made as the patient numbers available for download were different from the number of patients used in the Lawrence et al. [2014] main paper and from the number of patients used for this plot in the supplementary material (which is also different from the number stated in the main paper).

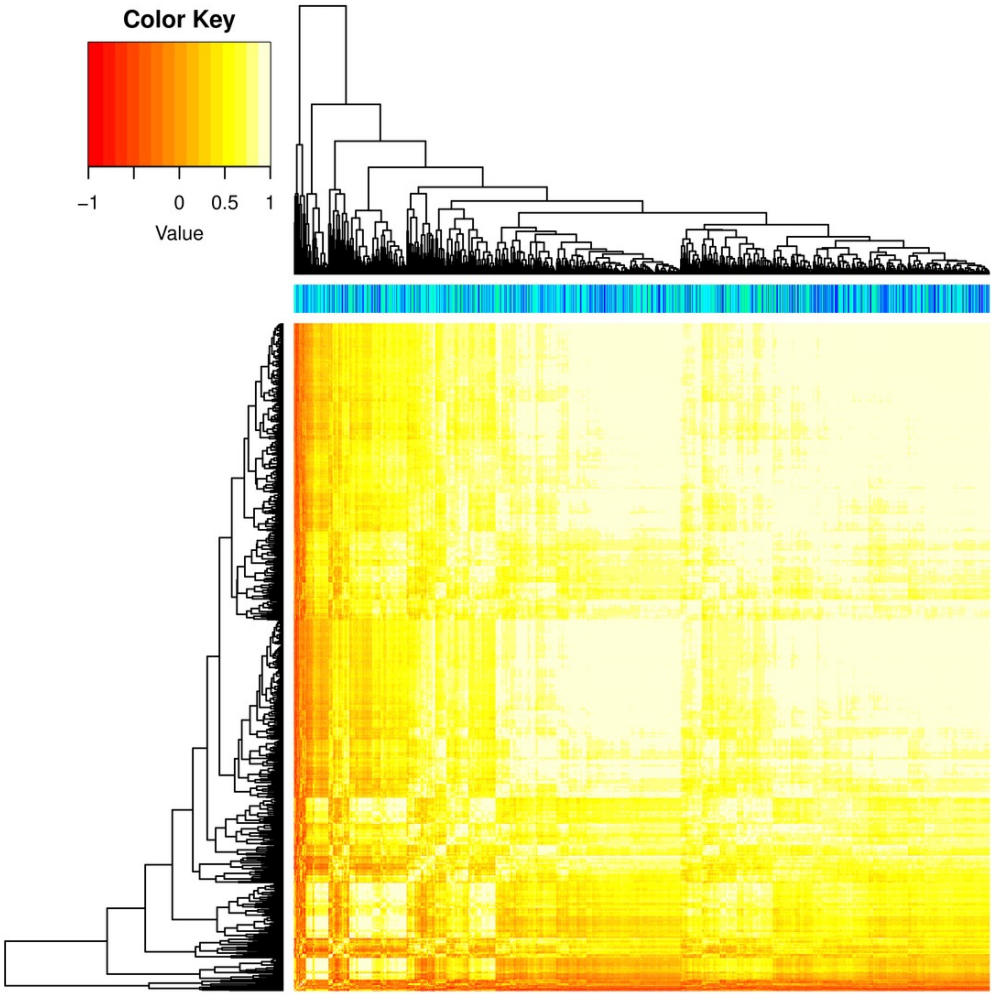


FIGURE 3.13: **Mutation spectra hierarchical cluster tree and heatmap over whole TCGA dataset.** This plot shows how all 1,005 patients in the TCGA dataset cluster by their mutation spectra, over all 17 different tumour types. Each tumour type is represented by a different colour in the horizontal panel across the top of the heatmap. The 17 different colour codes are well mixed across the dataset suggesting that, based on single nucleotide variants, these patients do not cluster by tumour type as might be expected. A colour key shows the scale of values within the heatmap, which are based on correlations between individuals, with a value of 1 indicating the highest identity between two patients.

TABLE 3.7: **Mutation profile proportions in Lawrence dataset.** For each of the 21 cancer types in the coding SSNVs Lawrence dataset of 4,712 patients, the mean proportion ± standard deviation (3dp) per patient has been tabulated for each of the six mutation classes for all coding SSNVs. The proportion has been calculated as a proportion of the total number of variants in that cancer type for that patient. For each patient the proportions were calculated and then the mean and standard deviation was taken over all patient proportions for each cancer type. The highest proportion for each cancer type is highlighted in blue.

Disease	Mutational signature					
	A⇒C/ T⇒G	A⇒G/ T⇒C	A⇒T/ T⇒A	C⇒A/ G⇒T	C⇒G/ G⇒C	C⇒T/ G⇒A
BLCA	0.023±0.022	0.079±0.051	0.029±0.023	0.111±0.066	0.249±0.100	0.509±0.083
BRCA	0.046±0.048	0.113±0.073	0.055±0.050	0.144±0.096	0.136±0.099	0.507±0.139
CARC	0.085±0.065	0.161±0.075	0.092±0.059	0.143±0.089	0.107±0.101	0.412±0.156
CLL	0.047±0.060	0.183±0.100	0.063±0.096	0.191±0.147	0.063±0.065	0.452±0.162
CRC	0.047±0.032	0.100±0.054	0.046±0.033	0.142±0.061	0.056±0.041	0.609±0.108
DLBCL	0.088±0.065	0.142±0.065	0.066±0.039	0.129±0.113	0.076±0.043	0.499±0.139
ESO	0.138±0.122	0.125±0.057	0.053±0.026	0.133±0.053	0.068±0.044	0.483±0.127
GBM	0.031±0.027	0.122±0.076	0.044±0.030	0.106±0.052	0.072±0.039	0.626±0.109
HNSC	0.031±0.030	0.110±0.071	0.059±0.063	0.153±0.089	0.156±0.091	0.492±0.145
KIRC	0.084±0.041	0.170±0.060	0.107±0.046	0.197±0.064	0.116±0.051	0.326±0.082
LAML	0.032±0.055	0.116±0.121	0.067±0.139	0.124±0.140	0.061±0.085	0.601±0.202
LUAD	0.036±0.043	0.087±0.047	0.083±0.057	0.365±0.148	0.123±0.059	0.306±0.138
LUSC	0.029±0.013	0.104±0.034	0.077±0.028	0.333±0.119	0.159±0.071	0.298±0.093
MED	0.038±0.086	0.106±0.122	0.041±0.068	0.205±0.173	0.063±0.123	0.547±0.205
MEL	0.023±0.021	0.054±0.043	0.030±0.020	0.051±0.053	0.031±0.042	0.810±0.148
MM	0.080±0.061	0.140±0.072	0.070±0.047	0.130±0.065	0.094±0.058	0.485±0.115
NB	0.051±0.164	0.059±0.068	0.052±0.065	0.427±0.241	0.100±0.134	0.311±0.208
OV	0.060±0.035	0.127±0.053	0.080±0.045	0.177±0.065	0.162±0.070	0.394±0.117
PRAD	0.053±0.059	0.124±0.118	0.055±0.056	0.154±0.118	0.087±0.083	0.526±0.178
RHAB	0.065±0.111	0.064±0.104	0.026±0.076	0.126±0.129	0.085±0.152	0.633±0.241
UCEC	0.037±0.031	0.114±0.069	0.032±0.026	0.146±0.068	0.078±0.075	0.592±0.115

It is clear from the pattern of mutation spectra across patients in Figure 3.14 that melanoma (MEL) has the highest median mutation frequency of all the tumour types, and rhabdoid tumour (RHAB) patients have the lowest mutation rates. These results support findings of higher mutation rates in cancers affected by environmental exposures such as UV radiation, and lower mutation rates in childhood cancers [Meyerson et al., 2010, Stratton, 2011]. Most of the cancers shown in Figure 3.14 have a more common intermediate frequency shown by the mutation frequency curves flattening out in the

TABLE 3.8: **Transition:transversion ratios in Lawrence data.** For each of the 21 cancer types in the coding SSNVs Lawrence dataset of 4,712 patients, the transition:transversion ratio (3dp) has been calculated by dividing the total number of transition mutations over all patients in a cancer type by the number of transversions for that cancer type. The number of patients in each subset has not been accounted for since that is not necessary when estimating ratios. The cancer type with the highest ratio has been highlighted in red.

Cancer type	TS:TV
BLCA	1.430
BRCA	1.364
CARC	1.324
CLL	1.596
CRC	2.029
DLBCL	2.653
ESO	1.479
GBM	2.882
HNSC	1.259
KIRC	0.962
LAML	2.593
LUAD	0.417
LUSC	0.700
MED	1.925
MEL	11.120
MM	1.611
NB	0.766
OV	0.956
PRAD	1.833
RHAB	1.527
UCEC	1.892

middle, with the high and low rates being rarer.

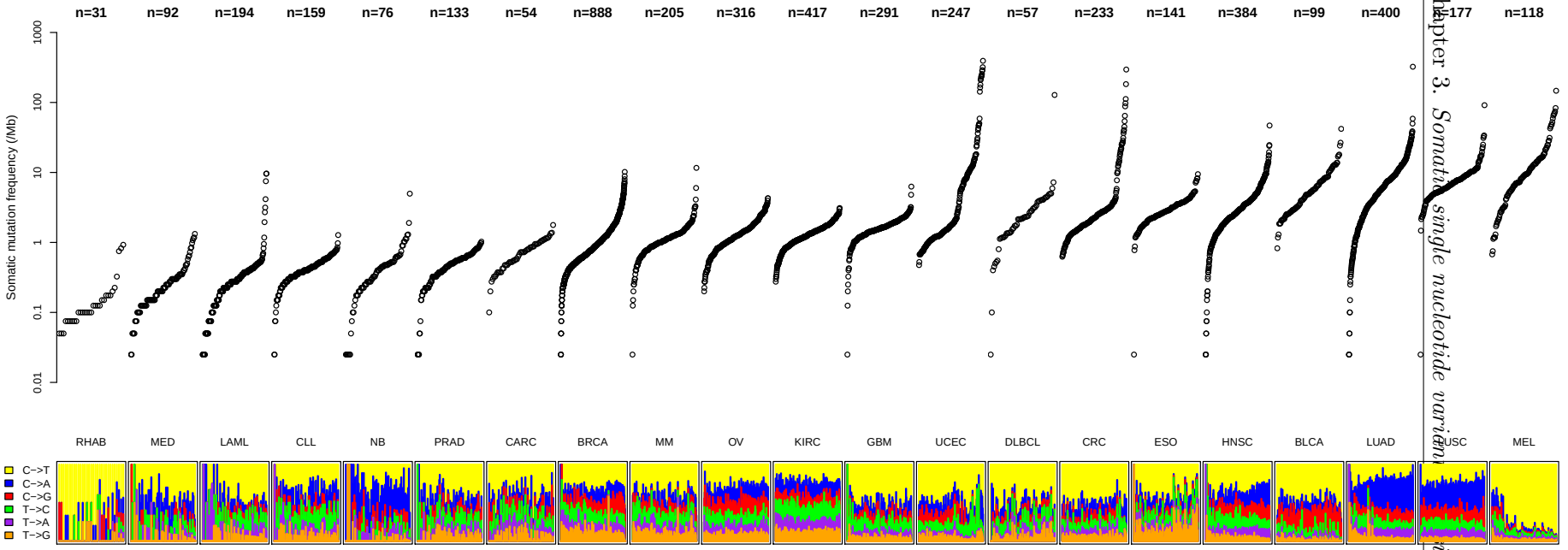


FIGURE 3.14: **Distribution of mutation rates and spectra across tumour types in the Lawrence dataset.** Using the set of cancer-specific coding SNVs, the mutation frequency and spectra has been plotted for each of the 21 tumour types over 4,712 patients. Each dot represents a single patient, with the y-axis value indicating the total frequency of somatic SNVs in the exome per Mb. Tumour types have been ordered by their median somatic mutation frequency in ascending order. Within each tumour type, individual patients have also been ordered in ascending somatic mutation frequency along the x-axis. The mutation rates have been log transformed before plotting to spread out the data points for clearer visualisation, but the annotated y-axis values are the actual frequencies. The lower panel shows the relative proportions of the six different possible base-pair substitutions, as indicated in the legend. Sample sizes (patient numbers for each tumour type) are shown along the top of the plot. None of the 291 Lawrence GBM patients were lost when filtering out the 16 patients with no coding SSNVs. *R code used to generate plot can be found in Supplementary Appendix C, re-implemented from scratch.*

The mutation spectra proportions have been plotted as a bubble plot in Figure 3.15, using all 4,712 patients in the Lawrence dataset. Of the 12 possible mutation changes, the highest proportions are C→T (G→A) transitions, as was seen in the TCGA dataset in Figure 3.11.

In Figure 3.14 it can be seen that the majority of MEL patients appear to have a very specific mutation profile with mostly C→T mutations (or G→A on the complementary strand), although the MEL patients with lower mutation frequencies have a much smaller proportion of this particular mutation class. The larger group of melanoma patients exhibit a spectrum indicative of UV-induced damage, since UV exposure is known to cause a signature of high mutation rates in melanoma with mostly C→T substitutions. To illustrate further how mutation spectra can vary within a particular tumour type, the 118 MEL patients in the Lawrence dataset were split into two distinct groups based on their mutation spectra. This was done using cluster analysis and plotted in a dendrogram in Figure 3.16. The red line shows where the data was split into two groups. In Figure 3.17, the two MEL groups from Figure 3.16 were plotted in two separate bubble plots. As can be seen the larger subset 'Group 2' has a far higher proportion of C→T changes than the smaller 'Group 1', which is also observed in Figure 3.14. 'Group 1' also has more C→T changes than any other change but it is less pronounced than in 'Group 2'. This much higher frequency of C→T transitions compared to that normally expected in human evolution is a signature of UV radiation [Pleasance et al., 2010a], which is known to contribute to the development of melanoma. This result suggests that the smaller subset of patients with much lower mutation rates may be affected by tumours that are not caused by UV-induced damage.

Approximately 50% of melanomas are known to harbour activating mutations in BRAF, which are known to be most common in patients whose tumours arise on skin without chronic sun-induced damage [Ascierto et al., 2012]. Among the BRAF mutations observed in melanoma, over 90% are at codon 600, and among these over 90% are a single nucleotide substitution resulting in a valine to glutamic acid change (BRAFV600E: nucleotide 1799 T→A (A→T on complementary strand); codon GTG→GAG; genomic

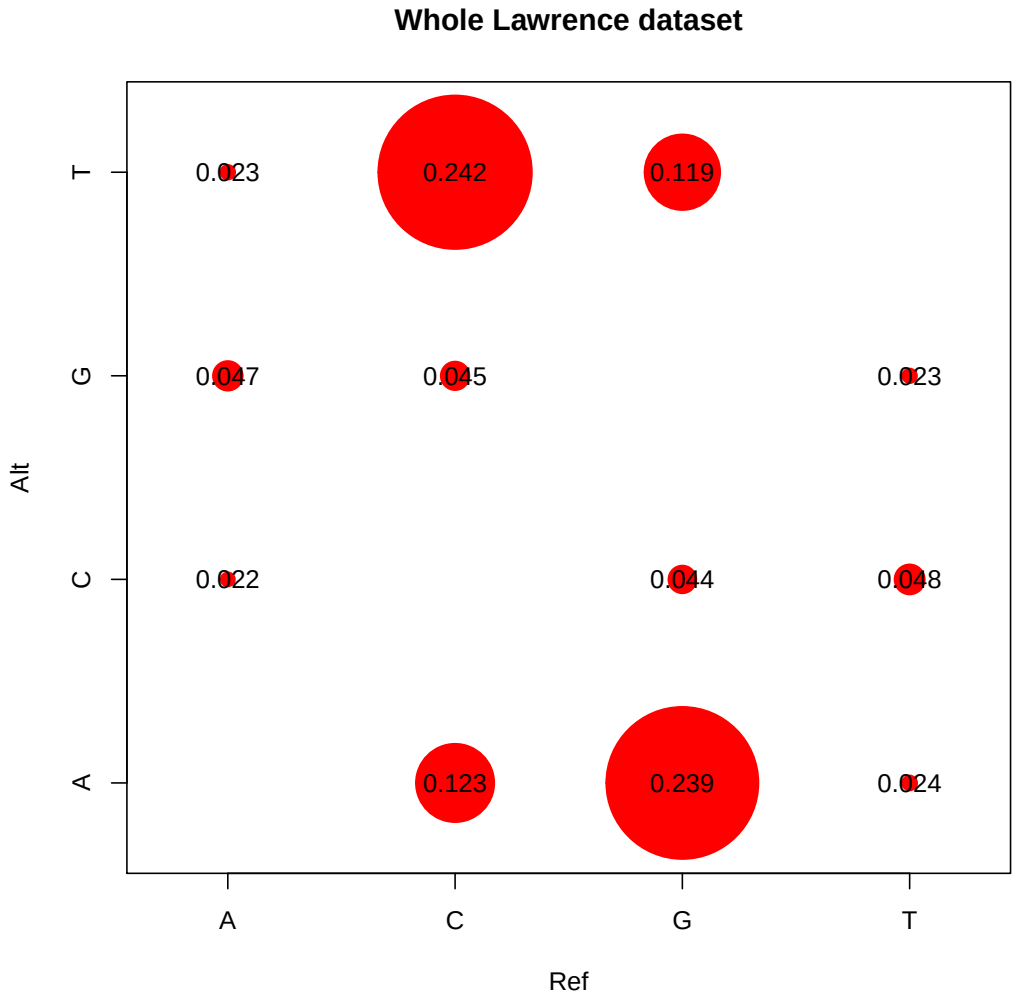


FIGURE 3.15: **Mutation spectra bubble plot for whole Lawrence dataset.** For whole coding SSNVs Lawrence dataset over all 21 tumour types and 4,712 patients, the proportions of the 12 different possible base-pair substitutions have been shown, with the four possible reference bases along the x-axis and the four possible alternate bases (that the reference base has mutated to in the cancer) along the y-axis. The size of the red circles represent the proportion of nucleotide substitutions from the total number of cancer-specific mutations. Raw counts over whole dataset have been used to calculate proportions, rather than mean values per patient. For each mutation class, the proportion was calculated by dividing the mutation count for that mutation class by the total number of mutations in the dataset.

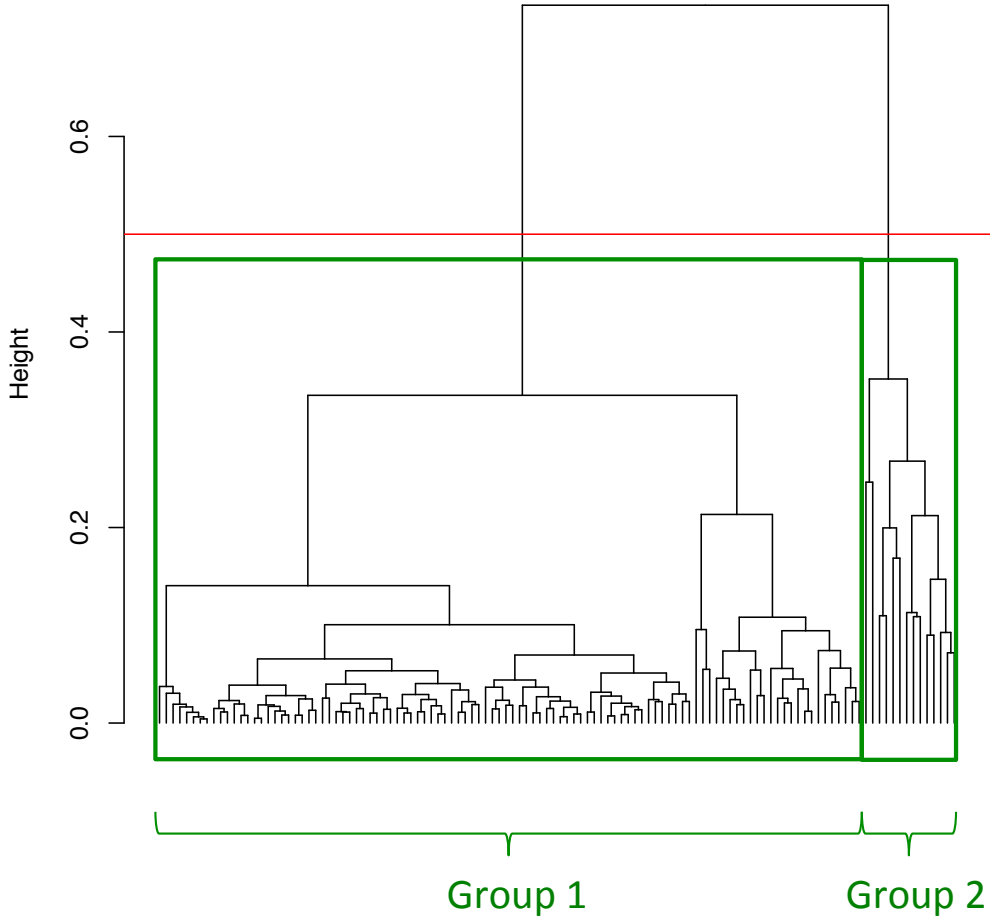
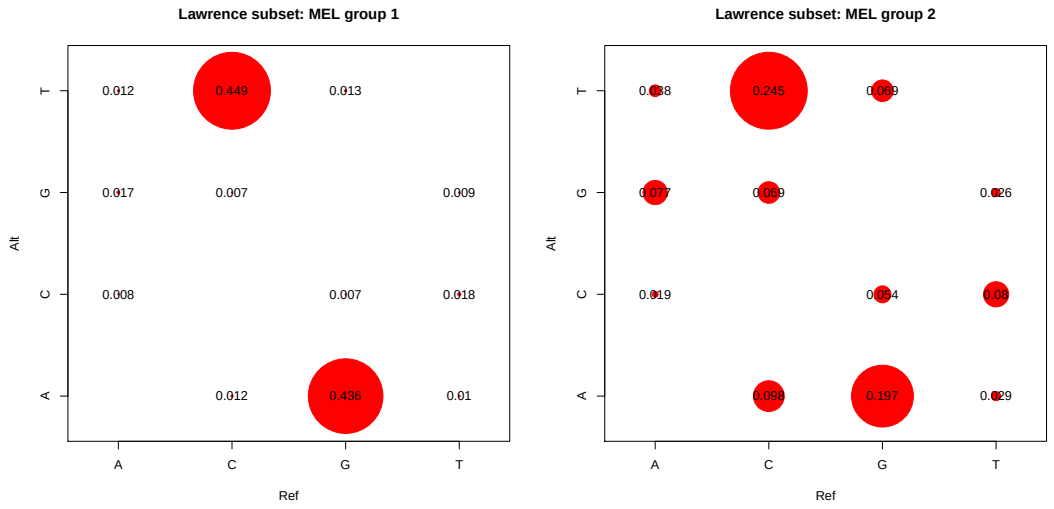


FIGURE 3.16: **Dengrogram of MEL patients in coding SSNVs Lawrence dataset clustered by single nucleotide variant spectra.** Dendrogram generated through cluster analysis shows how the 118 MEL patients in the Lawrence dataset cluster by mutation spectra. A horizontal red line has been drawn at the point where the data can be split into two groups ('Group1' on the left containing 104 patients and 'Group 2' on the right containing 14 patients) with distinctly different mutation spectra.



(A) MEL group 1 (B) MEL group 2

FIGURE 3.17: **Mutation spectra bubble plots comparison by mutation signature for two MEL sub-groups in Lawrence dataset.** For (A) all 104 patients in MEL "group 1" and (B) 14 patients in MEL "group 2" in the coding SSNVs Lawrence dataset of 4,712 patients, the proportions of the 12 different possible base-pair substitutions have been shown, with the reference allele along the x-axis and the alternate allele along the y-axis. The size of the red circles correspond to the size of the proportion. Raw counts over whole dataset have been used to calculate proportions, rather than mean values per patient. For each mutation class, the proportion was calculated by dividing the mutation count for that mutation class by the total number of mutations in the dataset.

position 140453136). It therefore may be expected that the smaller subset of patients with lower mutation rates are more commonly affected by BRAF mutations than the larger group of melanoma patients with mostly C→T mutations, and could explain the difference in mutation spectrum between the two groups. This was investigated by searching for enrichment of BRAF mutations in both groups. In 'Group 1' there are a total of 46,307 non-synonymous mutations (including 43,682 missense, 2,624 nonsense and 1 nonstop) spread over 12,255 genes. In 'Group 2' there are a total of 702 non-synonymous mutations (including 652 missense and 50 nonsense) spread over 637 genes. Most of these genes in both groups are hit by a single non-synonymous mutation, however the most commonly mutated gene in 'Group 1' is TTN hit 441 times, and in 'Group 2' is APC hit 8 times across the dataset. BRAF is the 11th most commonly hit gene in 'Group 1' with 66 non-synonymous mutations in 64 different patients, and the

12th most commonly hit gene in 'Group 2' with 3 non-synonymous mutations in 3 patients. Overall 69 of the 114 (61%) patients with melanoma have mutations in BRAF. Within 'Group 1' 64 of the 104 (62%) patients harbour mutations in BRAF, however within 'Group 2' only 3 of the 14 (21%) patients have non-synonymous mutations in BRAF. Of the 66 BRAF mutations in 'Group 1', 59 (89%) are the V600E mutation. Of the 3 BRAF mutations observed in 'Group 2', all (100%) are the V600E mutation. Overall, this suggests that BRAF mutations are not the source of the specific mutation profile in 'Group 2'.

As has been shown previously [Pleasance et al., 2010a,b], lung cancer and melanoma are known to have a specific mutational profile due to their exposure to tobacco smoke and UV radiation respectively. Therefore it might be expected that other cancer types also have a specific mutational profile, which has been investigated in Figure 3.18. However, this figure shows that the clustering of 4,712 patients containing coding SSNVs in the Lawrence dataset based on mutation spectra does not seem to correlate with the tissue of origin, which was also seen in Figure 3.13. However, it is difficult to see how exactly the tumour types are spread throughout the clusters, since there are so many patients in this dataset. The fact that patients of the same cancer type do not cluster together by mutational profile suggests that mutation spectra is not primarily dependent on the tissue of origin.

3.2.4.3 TCGA and Lawrence dataset comparison

The mutation spectra of two different cancers from each dataset have been compared in the heatmaps in Figure 3.19. GBM and OV have been compared from the TCGA dataset, and CRC and NB from the Lawrence dataset. GBM and OV show very little clustering by tumour type, suggesting that these two cancers do not have their own distinctive mutation spectra which distinguishes them from each other. This result is expected, as in Figure 3.9 GBM has some patients with a similar mutation profile to OV patients. The similarities between the mutation spectra of some of the GBM patients (with high mutation rates) and the OV patients in TCGA was the motivation for using

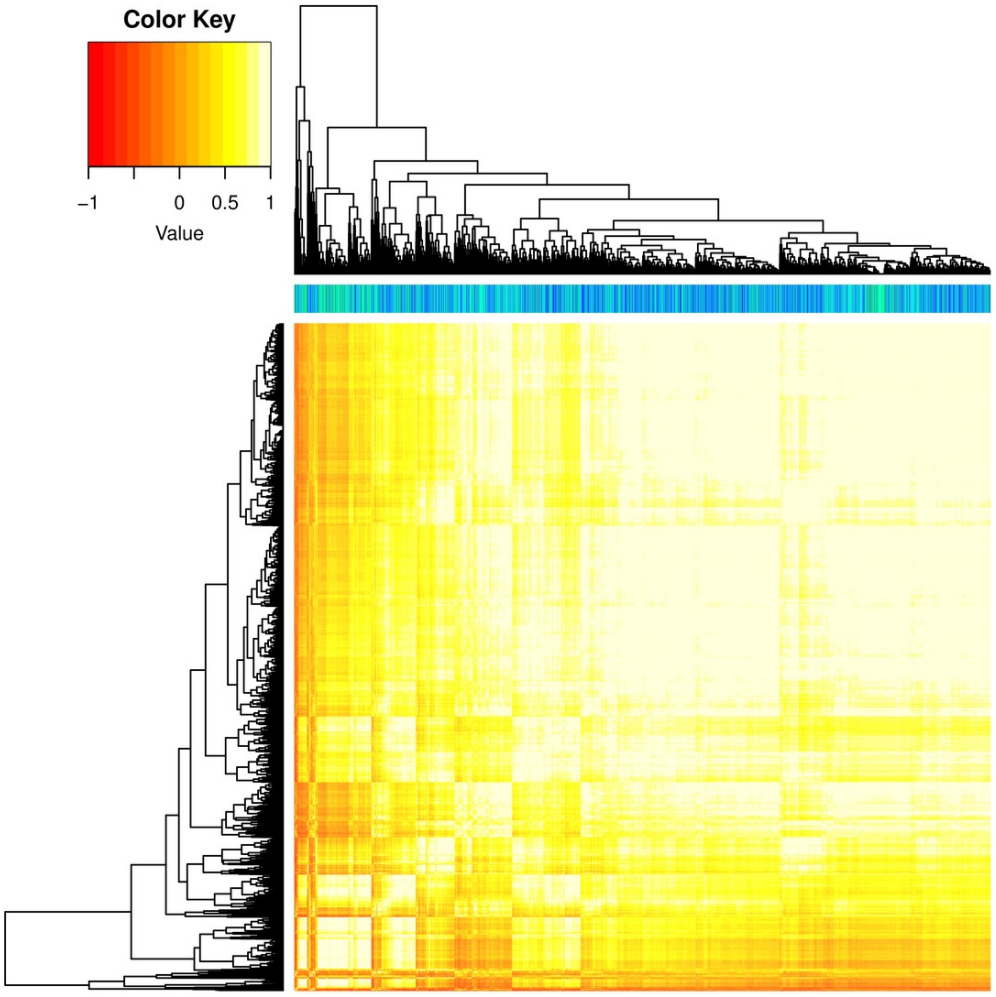
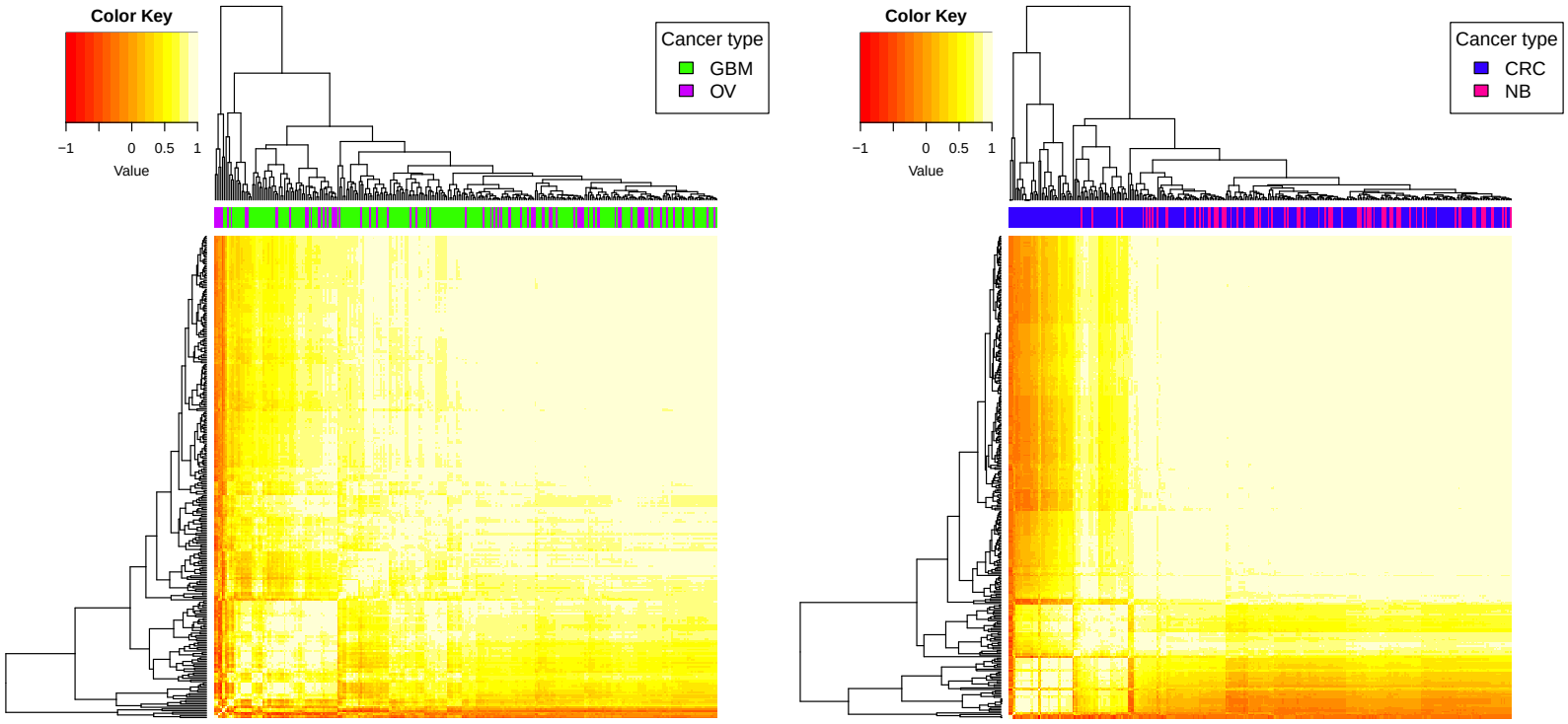


FIGURE 3.18: **Mutation spectra hierarchical cluster tree and heatmap over whole Lawrence dataset.** This plot shows how 4,712 patients in the coding SSNVs Lawrence dataset cluster by their mutation spectra, over all 21 different tumour types. Each tumour type is represented by a different colour in the horizontal panel across the top of the heatmap. The 21 different colour codes are well mixed across the dataset, suggesting that based on single nucleotide variants, these patients do not cluster by tumour type as might be expected. A colour key shows the scale of values within the heatmap, which are based on correlations between individuals, with a value of 1 indicating the highest identity between two patients.

these two cancers in this heatmap comparison. However, slightly more clustering of mutation profiles by tumour type is observed in the heatmap comparing CRC and NB patients. A sub-group of CRC appears to have a specific mutational profile shown by the cluster of blue in the colour-coded panel representing CRC patients. Colorectal cancer is known to have a specific mutational spectrum with elevated rates of single nucleotide substitutions at polynucleotide tracts caused by mutations in DNA mismatch repair genes, supporting this finding.

Combined with results from Figure 3.13 and Figure 3.18 it can be concluded that mutation spectra does not necessarily cluster by tissue of origin. However the varied results suggest that mutational profile as well as tumour type should be corrected for before performing evolutionary analysis, by partitioning the data according to these factors.



(A) GBM and OV from TCGA dataset

(B) CRC and NB from Lawrence dataset

FIGURE 3.19: **Mutation spectra hierarchical cluster tree and heatmap comparisons between two different tumour types.** Patients were clustered by their mutation spectra for (A) GBM and OV in the TCGA dataset and for (B) CRC and NB in the coding SSNVs Lawrence dataset. In the comparison between 208 GBM patients and 75 OV patients, there appears to be little clustering by tumour type, as can be seen by the colour-coded panel in (A). However, in the comparison between 233 CRC patients and 76 NB patients in the Lawrence dataset, the colour panel representing the tumour types shows slightly more clustering by tumour type, as some of the patients that have been clustered together by mutation spectra in (B) are also all CRC patients, suggesting that this tumour type may have a specific mutation spectra in a subset of patients. Colour keys show the scale of values within the heatmaps, which are based on correlations between individuals, with a value of 1 indicating the highest identity between two patients. Each tumour type is represented by a different colour in the horizontal panel across the top of the heatmap, for which there is also a key for each heatmap.

3.3 Discussion

3.3.1 Transition mutations occur at higher rate than transversion mutations in most cancers

The characterisation of somatic SNVs over the whole TCGA dataset of 1,005 patients and the coding Lawrence dataset of 4,712 patients has revealed that most tumour types exhibit an increased rate of transitions compared to transversions, with the exception of glioblastoma multiforme (GBM), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC) and ovarian serous cystadenocarcinoma (OV) in the TCGA dataset, and KIRC, LUAD, LUSC, neuroblastoma (NB) and OV in the Lawrence dataset. Analyses of neutrally evolving genomic sequences reveal that transition mutations occur at significantly higher rates than transversions. This is also reflected in cancer sequencing studies [Rubin and Green, 2009], so the results in this analysis are expected. In the cases of LUAD and LUSC, the increased transition mutational pattern is known to be perturbed by environmental exposures that result in many more C→A/G→T transversions than is normally seen in the human genome or in other cancer types, induced by tobacco carcinogens [Stratton, 2011], again an expected result. The increased transversion rates observed in other cancer types such as GBM, KIRC, NB and OV could therefore indicate interesting specific mutational processes underlying the initiation and progression of these cancers.

3.3.2 Mutation spectra varies between and within tumour types

The results here show that there is a great deal of mutational heterogeneity both between tumour types and between samples within tumour types. For example, the six substitution class classification system has revealed that mutation spectra does not necessarily correlate with tissue of origin, suggesting that the mutational profile of a cancer is also an important consideration when attempting to determine the underlying mutational processes of cancers.

Based on this finding, it was concluded that both tissue of origin and mutational spectrum should be controlled for before driver mutation detection analyses, by grouping data by tumour type and also by partitioning data based on mutation spectra. Lawrence et al. [2014] have only addressed how the tissue of origin affects the genes hit by driver mutations, so by also incorporating mutational signatures this work can be built on to provide a more comprehensive understanding of the mechanisms underlying cancer progression and detect driver genes that would otherwise remain undetected due to a lack of power. This work has been carried out in Chapter 6.

3.3.3 Increased rate of INDELs in subset of GBM patients

During the comparison of TCGA and Lawrence calling pipelines using the 701 patients that have had their mutations called by both pipelines, a subset of 16 GBM patients were found to have much higher mutation rates in the TCGA dataset compared to the Lawrence dataset. These outliers were an artefact of the data that on further investigation, as well as highlighting differences between the two calling pipelines, represented a mutation spectrum of a very high rate of called INDELs in these patients. It is the high rate of INDELs in these patients that are thought to have caused the high rate of miss-called SNVs in the TCGA pipeline, as SNVs are more difficult to call next to INDELs. It is possible that Lawrence have accounted for this in their pipeline by removing SNVs around INDELs, thus avoiding false-positives. This signature is however not representative of all GBM patients, suggesting the presence of a specific mutational spectrum characterised by INDELs in a subset of GBM patients. It would be interesting to further investigate the aetiology underlying this spectrum, which is carried out to some extent in Chapter 7, where it is suggested that these called INDELs are actually miss-called larger genomic rearrangements.

3.3.4 Caveats

3.3.4.1 A→G/T→C mutational asymmetry

Single nucleotide mutations have been categorised into six classes in this analysis, assuming that complementary mutations (e.g. A→G and T→C on the complementary strand) occur at the same rate. However it has been shown that A→G substitution rates are elevated in all cancer types compared to the complementary rate of T→C mutations on the coding strand, suggesting that mutation patterns within genes are influenced by gene expression. The asymmetry has been found to be statistically significant in breast cancer [Rubin and Green, 2009]. Therefore, the assumption of mutational symmetry may not always be accurate and may justify using 12 single nucleotide mutation classes for more refined mutational profiles, to improve the classification of cancers by their specific mutational profile, especially in cancers where gene expression plays a significant role. However, many cancer studies have used just six classes of base substitution based on strand symmetry, rather than 12 [Alexandrov et al., 2013, Kandoth et al., 2013, Lawrence et al., 2013], so this is considered common practice.

3.3.4.2 Incorporating sequence context into mutational spectra classifications

In this analysis, only the numbers of each class of mutation have been used to create mutation profiles. However, it would be beneficial to further define mutational signatures by incorporating the sequence context of each mutation into the set of features. For example, in Alexandrov et al. [2013] information on the bases immediately 5' and 3' to each mutated base was incorporated on top of their six classes of base substitution (C→A, C→G, C→T, T→A, T→C, T→G) to provide 96 possible substitution mutations in their classification. This classification system would be particularly useful in distinguishing mutational signatures that cause the same substitutions but in different sequence contexts. For example, this 96 substitution classification would help aid the distinction between C→T mutations at NpCpG trinucleotides thought to be related to

the relatively elevated rate of spontaneous deamination of 5-methyl-cytosine [Alexandrov et al., 2013], and C→T mutations at TpCpN trinucleotides caused by over activity of members of the APOBEC family of cytidine deaminases responsible for converting cytidine to uracil and thought to be induced by certain classes of viruses [Alexandrov et al., 2013, Lawrence et al., 2013].

3.3.4.3 Increased rate of called SNVs around INDELs in TCGA pipeline

As previously mentioned, the TCGA pipeline has been shown to over-call SNVs that are proximal to INDELs, illustrated by the 16 GBM patients with much higher SNVs called by the TCGA pipeline compared to the Lawrence pipeline. This shows that GATK used by the TCGA pipeline to call SNVs and INDELs is not optimal at calling SNVs around INDELs compared to the pipeline used in Lawrence et al. [2014]. The TCGA data is also not effectively set up for detecting INDELs reliably; whole-genome sequencing data would be better suited to the identification of INDELs [Meyerson et al., 2010]. Furthermore TCGA is a smaller dataset than Lawrence so there is less power to detect INDELs in the TCGA dataset. However, it is not known if Lawrence had the same problems in variant calling and simply excluded SNPs proximal to INDELs after variant identification, in which case an additional filtering step could be undertaken in this analysis to remove suspected false-positive SNVs near INDELs.

It would be useful to also investigate INDELs in these 16 GBM patients in the Lawrence dataset as well as in the TCGA dataset, since INDEL mutation data was also available for download from the Lawrence dataset.

3.3.4.4 Lawrence filtering steps not well documented

Lawrence et al. [2014] have not been explicit about how their somatic SNVs were filtered prior to analysis. For example, it was not mentioned if homozygous mutations were removed from the Lawrence data as they were in the TCGA data, so it was assumed that they were not excluded. It was also not clear what significance threshold was used

for the highly significantly mutated genes in the MutSig analysis, so this had to be calculated manually.

There was also confusion over the mutation type classifications, for example both silent and synonymous terminology was adopted in the Lawrence mutation data. However, it was assumed that these were all of synonymous consequence, since coding mutations were referred to as only missense or synonymous in the Lawrence et al. [2014] paper, and silent mutations were not mentioned.

Additionally, there is a discrepancy between the numbers of SNVs reported in the Lawrence et al. [2014] paper and the number of SNVs analysed in the data available for download, which is not explained anywhere in the documentation.

Finally, 37 GBM patients from the TCGA were not included in the Lawrence dataset despite being available for download at the time of the [Lawrence et al., 2014] study. The reason for this exclusion is not known, however it is suspected that it is due to the very high mutation rates exhibited by these patients. It would be helpful to investigate these patients further by measuring the INDEL rates, which could be elevated and hence causing the high SNV mutation frequency. It is possible that these patients were removed from the Lawrence dataset due to a high rate of INDELs and a resulting high rate of false-positive SNVs near INDELs. However, if this was the case then the 16 outlier GBM patients present in both datasets with high INDELs found in TCGA would also have been removed from the Lawrence dataset, which they have not been. Further filtering information is needed to further understand these anomalies.

3.3.4.5 TCGA-Lawrence comparison

It was difficult to make direct comparisons between the TCGA and Lawrence datasets in order to access the performances of the different pipelines, due to the fact that different cancer types and patient numbers were covered by each of the datasets. However, there was an overlap of 701 patients between the two datasets which made a direct comparison possible. Further filtering of the Lawrence data in this analysis has been

performed to attempt to make the comparison as fair as possible by only using the coding mutations from the Lawrence dataset as has been used in the TCGA dataset, and only using single nucleotide mutations in both datasets.

3.4 Methods

For this chapter the filtered set of cancer-specific heterozygous non-synonymous and synonymous SNVs from TCGA were used. However, in this set the same mutation in a patient is reported multiple times in cases where there are multiple possible transcripts for the gene that it appears in. This is due to using the 'consequence' table in the *tcga_pair_exome* database to create the filtered set of heterozygous cancer-specific TCGA SNVs. The consequence of a mutation is vital for the evolutionary analysis later on, and for that the consequence of the longest transcript was used. However this set has been further filtered to contain each mutation in a patient once using the most severe consequence. The Lawrence dataset as provided had already been filtered to contain each mutation only once, however it was not documented which transcript was used for each gene to give consequence information for each mutation.

3.4.1 Summary variant statistics on a per patient basis

The tabulated mean mutation frequencies were calculated in R by dividing the total mutation count for each tumour type by the number of patients in that tumour type subset to give a mean count per patient for each tumour type, rounded to the nearest whole number. Histograms were generated in R using raw mutation counts for each patient. Mutations were only counted once if they occurred in overlapping genes. This was done for both TCGA and Lawrence datasets.

3.4.2 SSNVs on a per gene basis

The total coding mutation count for each gene over the whole dataset was divided by the number of patients in each dataset, so by 1005 in TCGA and 4712 in Lawrence, to get the mean number of mutations per patient for each gene.

For the gene length normalised plots, the mean counts per patient were divided by the CDS length of the transcript with the longest CDS length for each gene (not necessarily the CDS length of the longest transcript) to give a mutation count per nucleotide. A list of Ensembl CDS lengths was obtained from Biomart [Kasprzyk, 2011]. Perl was used to find the length of the longest CDS for each gene.

A single mutation was counted twice in overlapping genes for this gene-based analysis.

3.4.3 Patients overlapping datasets

For the 701 coding patients that are present in both TCGA and Lawrence datasets, a comparison was made between their mutation calls in TCGA and in Lawrence, by plotting the mutation counts in the TCGA dataset against the corresponding counts in the Lawrence dataset for each patient in R using the *plot* function.

To test the association/ correlation between the paired samples in Figure 3.4, a correlation coefficient and p-value was calculated using *cor.test* in R. This test is given two vectors of numbers: the mutation count for the 701 patients in TCGA and the mutation counts for the 701 patients in the Lawrence dataset. The default method is Pearson's product-moment correlation, which was used here. The default alternative hypothesis is two-sided, also used here, which means the alternative hypothesis is that the true correlation is not equal to 0 and the null hypothesis (rival hypothesis) is that the correlation is 0 (no correlation).

3.4.3.1 INDELs in GBM outlier patients

The 16 GBM patients with elevated mutation rates observed in the TCGA dataset compared to the Lawrence dataset were investigated further by mining the mySQL database containing TCGA mutations to look for the presence of heterozygous cancer-specific INDELs in these patients. To create Figure 3.8, the total number of INDELs counted across each of the six groups of patients were divided by the number of patients present in that subset in order to obtain a mean value per patient. In mining the database, the same patient mutation appeared multiple times due to multiple possible transcripts that the mutation could affect, therefore duplicate mutations for the same patient were removed before calculating INDEL counts, and mutations in overlapping genes were counted only once. Two patients in the database were not used in the filtered set of cancer-specific SNVs as they contained no cancer-specific SNVs, so for consistency these two patients were also removed from the INDEL datasets. The 'consequence' table from *tcga_pair_exome* was used, meaning that only coding INDELs were included in this analysis, ignoring all INDELs that occur in introns. The cancer-specific SNVs are also only coding variants.

3.4.4 Mutation spectra

For Table 3.5 and Table 3.7, for each patient the mutation profile proportions were calculated by dividing the number of mutations for each signature by the total number of mutations in that patient. Then over all patients, the proportions for each signature were summed and then divided by the number of patients to get the mean proportion per patient. This was done for each cancer type subset to produce the tables. Standard deviation was also calculated, all in R.

For Figure 3.9 and Figure 3.14, the mutation count for each patient was converted into a mutation rate per Mb. This was done by dividing the mutation count by 40,000,000 to get the rate per base, and then multiplying by 1,000,000 to get the rate per Mb. The mutation frequency over the whole exome was divided by 40,000,000 because 40

Mb (∼1.3% of the whole genome) is the typical length of the exome target region from an Illumina exome capture panel. The size of the target region using Illumina would be 62 Mb if flanking regions relative to each bait were included, which they are in Illumina, but is equivalent to 40Mb (for coding regions only) when compared to other kits. Patients were then ranked in ascending order of their mutation frequency and plotted as a distribution, and split by cancer type, to show the range of mutation frequency within cancer types. In the same plot, the relative proportions of each of the six SNV classes were calculated for each patient, and plotted as a cumulative bar chart to show the variation of mutation spectra within cancer types relative to the mutation frequency. This was done for both the TCGA dataset, and also separately for the Lawrence dataset.

Bubble plots (Figure 3.11, Figure 3.12, Figure 3.15, Figure 3.17) were also generated in R, using the symbols function to create circles corresponding to the size of the data points representing proportions.

For the hierarchical cluster tree heatmaps in Figure 3.13, Figure 3.18 and Figure 3.19, a directional substitution matrix (non-time-reversible, e.g. considering C→T as distinct from T→C) was created before cluster analysis in R. Whereas in PAML, the evolutionary model described later on, uses time-reversible (C↔T), which is a limitation of the model.

Chapter 4

Preliminary gene-based evolutionary analysis: Detecting selection on whole TCGA dataset

4.1 Introduction

The purpose of this evolutionary analysis was to identify the driver mutations amongst the set of candidate driver mutations found using the data processing pipeline (described in Chapter 2) and distinguish them from the inconsequential passenger mutations. This was done using measures of selection in the codeml program of PAML, to look specifically for signals of positive selection in genes indicating the presence of driver mutations. The aim was to find genes harbouring driver mutations and hence identify the genic drivers of cancer.

Evolutionary analysis was carried out in PAML, using the filtered, heterozygous, cancer-specific TCGA SNVs identified and processed in Chapter 2 and described in Chapter 3. The numbers used in this preliminary evolutionary analysis are smaller than the patient numbers described previously however, since this analysis was carried out on a

preliminary set containing just 998 patients before a further 7 patients were added to the dataset.

Analysis was carried out on a per-gene basis, over all 17 cancer types in the dataset. This chapter describes the prototype for more refined subsequent evolutionary analyses.

4.2 Results

4.2.1 Screen for positive selection in genes

Gene-based analysis was carried out in PAML using the codeml program, to show which genes driver mutations are preferentially found in.

For each gene over the whole TCGA dataset of 998 patients, an omega ratio was estimated as a measure of selection. An omega value >1 is evidence for positive selection, suggesting that the gene harbours driver mutations in cancer, because the non-synonymous substitution rate is greater than the synonymous substitution rate. Conversely, an omega <1 is indicative of negative selection as the mode of selection, as the non-synonymous mutations are being removed from the population. An omega value of 1 suggests neutrality, so the non-synonymous substitution rate is equal to the synonymous substitution rate. It is genes with omega values larger than 1 that are of interest, since it is these genes that are under positive selection and are therefore likely to contain cancer driver mutations.

A p-value was also calculated for each gene using the log likelihood values estimated from the codeml models, which was used to calculate a FDR value, to increase the confidence in the omega estimate for each gene.

The omega and FDR values have been plotted for each gene in Figure 4.1. FDR is plotted along the y-axis representing the significance of the positive selection detected, and the x-axis denotes the omega as a measure of the magnitude of selection.

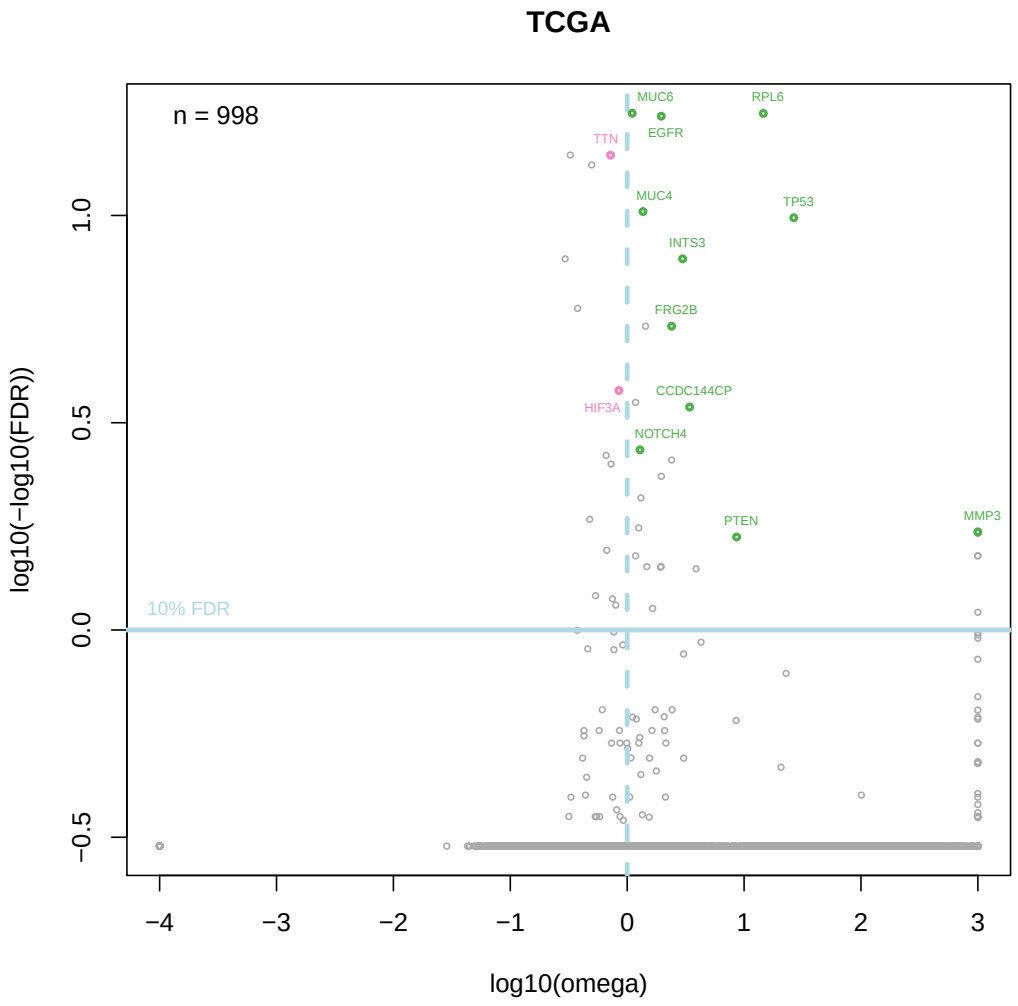


FIGURE 4.1: **Preliminary TCGA gene-based omega analysis in PAML.** This plot shows the results obtained from gene-based analysis in PAML on the whole TCGA dataset consisting of 998 patients over 17 different cancer types. The omega estimates have been plotted along the x-axis (log transformed) and FDR values along the y-axis (double log transformed). Each dot represents a gene. The blue vertical dashed line indicates point of selective neutrality (omega = 1), so genes under positive selection will fall on the right hand side of the vertical dashed line and some of these have been highlighted in green. The blue horizontal line represents the 10% false discovery rate threshold, which has been used as the significance threshold. Therefore all genes above this line are significantly hit by driver mutations. Genes highlighted in pink are those with significant FDR values supporting positive selection but omega values supporting purifying selection.

The blue vertical dotted line signifies an omega of 1 (log10(omega)=0), corresponding to neutral selection, so all genes to the right of this line have an omega ratio greater than 1 indicating that they are under positive selection. All genes to the left of this line have more synonymous mutations than non-synonymous mutations and hence are assumed to be under purifying selection. The blue horizontal line denotes the 0.1 false-discovery rate threshold (log10(-log10(FDR))=0), so no more than 10% of the genes above this line are false-positives. FDR of 0.1 has been used as the significance threshold in this analysis, so genes with a FDR<0.1 (above the horizontal blue line in Figure 4.1) are considered to be significantly mutated. The top right quadrant of the graph is therefore where we would expect to find candidate driver genes undergoing strong signals of positive selection with significant p-values and omega ratios indicative of positive selection, and some of these have been highlighted in green and annotated with their gene name. Genes highlighted in pink are those with significant p-values, but with omega values suggestive of purifying selection as the mode of selection acting on these genes.

The vertical line of genes along the far right-hand side of the plot are those with no synonymous mutations occurring (e.g. MMP3), so although they may have significant p-values the omega ratios are effectively infinity because the non-synonymous rate is divided by 0 in these cases and dividing by 0 gives a result of infinity. PAML automatically sets these omega values to "1000" (log10(1000)=3). These genes contain only the functionally important non-synonymous mutations.

In this preliminary TCGA dataset containing 998 patients, 25 genes have been found to be under strong signals of positive selection and therefore significantly mutated in cancer by driver mutations. These genes are listed in Table 4.1 with their omega and p-values as well as their gene descriptions. This table only contains the genes in the top right quadrant of Figure 4.1, so that is genes with both an omega suggestive of positive selection (omega>1) and a FDR value significantly supporting evidence for positive selection (FDR<0.1).

TABLE 4.1: **Ranked list of the 25 significantly mutated genes in TCGA whole-dataset analysis.** This tables shows the 25 genes found to be significantly hit by driver mutations in the gene-based analysis in PAML on the whole TCGA dataset of 998 patients over 17 different cancer types. These genes have been classed as significant based on the criteria that they have a FDR<0.1 and an omega>1, supporting evidence for positive selection acting in these genes in cancer.

Gene	Omega	P-value	Description
MUC6	1.10	1.73e-22	mucin 6, oligomeric mucus/gel-forming
RPL6	14.59	3.57e-22	ribosomal protein L6
EGFR	1.96	1.00e-21	epidermal growth factor receptor
MUC4	1.36	3.20e-14	mucin 4, cell surface associated
TP53	26.64	8.04e-14	tumor protein p53
INTS3	2.98	9.82e-12	integrator complex subunit 3
FRG2B	2.40	3.74e-09	FSHD region gene 2 family, member B
ANKRD36C	1.43	3.89e-09	ankyrin repeat domain 36C
DEPDC5	1.18	3.29e-07	DEP domain containing 5
CCDC144CP	3.42	4.30e-07	coiled-coil domain containing 144C, pseudogene
NOTCH4	1.29	2.46e-06	notch 4
ZSWIM8	2.40	3.88e-06	zinc finger, SWIM-type containing 8
ACKR4	1.96	7.15e-06	atypical chemokine receptor 4
DDX11	1.31	1.38e-05	DEAD/H (Asp-Glu-Ala-Asp/His) box helicase 11
GPATCH8	1.25	3.14e-05	G patch domain containing 8
MMP3	999.00	3.58e-05	matrix metallopeptidase 3 (stromelysin 1, progelatinase)
ANXA7	999.00	3.85e-05	annexin A7
PTEN	8.64	4.32e-05	phosphatase and tensin homolog
CYP27B1	999.00	7.09e-05	cytochrome P450, family 27, subfamily B, polypeptide 1
HNRNPC	1.18	7.18e-05	heterogeneous nuclear ribonucleoprotein C (C1/C2)
RPTN	999.00	7.28e-05	repetin
HNRNPCL1	1.47	9.49e-05	heterogeneous nuclear ribonucleoprotein C-like 1
POTEF	1.92	9.88e-05	POTE ankyrin domain family, member F
CEP164	3.89	1.05e-04	centrosomal protein 164kDa
DDX6	1.65	2.21e-04	DEAD (Asp-Glu-Ala-Asp) box helicase 6

4.2.2 Power to detect known cancer genes

These results demonstrate that this omega analysis has sufficient power to detect signals of positive selection in well-known cancer driver genes, for example the tumour suppressor gene TP53. EGFR and PTEN have also previously been found to be implicated in cancer and have been highlighted in green in Figure 4.1.

TP53 (tumour protein p53) encodes a regulator of the cell cycle machinery that can suppress the growth of cancer cells as well as inhibit cell transformation [Hollstein et al., 1994]. TP53 is the most commonly mutated gene in human cancer [Kandoth et al., 2013], inactivated by mutations that alter or obliterate normal TP53 function. As seen with other tumour suppressors, frameshift or nonsense mutations result in the loss of TP53 protein expression in some cases. However most often it is missense mutations that lead to a loss of TP53 wild-type activity [Muller and Vousden, 2014].

EGFR (epidermal growth factor receptor) is a growth factor receptor that induces cell differentiation and proliferation upon activation through the binding of one of its ligands. The receptor is located at the cell surface, where the binding of a ligand activates tyrosine kinase activity in the intracellular region of the receptor. This tyrosine kinase phosphorylates a number of intracellular substrates that activates pathways leading to cell growth, DNA synthesis and the expression of other oncogenes [Voldborg et al., 1997]. Activating mutations in EGFR have been found in 1520% of lung adenocarcinomas, and have been targeted by tyrosine kinase inhibitors such as gefitinib and erlotinib in the treatment of high-stage lung adenocarcinomas [Siegelin and Borczuk, 2014].

PTEN (phosphatase and tensin homolog) is a tumour suppressor gene that is deleted or mutated in a variety of human cancers including prostate, breast, brain, endometrial, lung, and ovarian cancers [Dong, 2001, Li et al., 1997]. The lipid phosphatase activity of PTEN is important for its tumour suppressor function, which operates by negatively regulating the PI3KAKTmTOR pathway [Hollander et al., 2011].

4.2.3 Candidate cancer genes

RPL6 has shown up as being very significant in this whole-dataset analysis, with a significant p-value of 3.57e-22 and a high omega ratio of 14.59, both strongly supporting positive selection acting on this gene in cancer. This gene is a ribosomal gene and has not been causally implicated in cancer [Futreal et al., 2004], however it has been found to be overexpressed in multidrug-resistant gastric cancer cells [Du et al., 2005].

4.2.4 Significant results for likely common false positives

TTN, MUC6 and MUC4 are all large genes that often show up as false-positives in cancer studies [Lawrence et al., 2013]. This is due to their size and subsequently higher number of mutations. However, in this evolutionary analysis, the size of the gene has been accounted for with the use of a synonymous substitution rate as a neutral proxy. All three of these genes are highlighted in Figure 4.1. TTN is highlighted in pink as it has an omega value suggestive of negative selection, and MUC4 and MUC6 both have been highlighted in green, as they both have an omega ratio indicative of positive selection. All three genes have significant p-values, which means there is strong evidence supporting positive selection. However it is suspected that these genes are false-positives.

4.2.5 Confounding signals of positive and negative selection within genes

Interesting genes are highlighted in pink in Figure 4.1 as they have an omega suggestive of negative selection (omega<1) but a significant p-value (strongly supporting good fit for model of positive selection). This could suggest that both positive and purifying selection is occurring in the same gene, but in different regions. Both TTN and HIF3A are genes showing evidence for undergoing both positive and purifying selection, with significant p-values and omega values close to 0 (indicating neutral selection). This

suggests that these genes could be undergoing different modes of selection within the same gene at different sites.

TTN often comes up as a false-positive in this type of analysis due to its size, however HIF3A could be a good candidate for sub-region analysis in which the gene is partitioned into functional domains and those domains are analysed separately to prevent competing signals confounding analysis. The purpose of this type of analysis is to identify the functional region that positive selection is targeting within a gene to further understand the mechanism involved in the progression of the cancer. HIF3A (hypoxia-inducible factor 3) encodes a protein that regulates many adaptive responses to low oxygen tension, and it is known that tumour hypoxia is a classical feature of cancer [Masson and Ratcliffe, 2014]. Therefore HIF3A could be a target for cancer driver mutations, but maybe has previously been missed in this type of analysis due to the fact that only a specific domain is subject to positive selection and maybe other parts of the gene are under purifying selection.

4.3 Discussion

Evolutionary analysis in PAML on a whole dataset has shown that PAML has sufficient power to detect well-known cancer genes such as TP53 and EGFR, showing proof of concept. Other genes such as MUC6 and MUC4 often show up as false-positives, so more work is needed to further understand these genes, and why this analysis has not successfully accounted for the size of the gene assuming that is the reason for the aberrant results. Genes undergoing purifying selection but with significant p-values such as TTN and HIF3A also require further analysis, to ascertain whether or not different types of selection are acting on the same gene, especially in HIF3A which is not normally found as a false positive. These genes are motivation for sub-gene analysis, to understand if selection is acting on just a specific part of these genes and is being confounded by other modes of selection simultaneously acting.

This analysis has also uncovered a potential candidate gene suspected to be implicated in cancer, RPL6, which is not already known to be a driver cancer gene. Many other genes that have not yet been associated with cancer have also been revealed. However, further work and validation is needed to conclude the role of these genes in cancer.

To further increase the confidence of results, the maximum likelihood estimated parameters could be used in the simulation of neutral evolution, to show what results would be expected under a random distribution of the observed mutation profile.

4.4 Methods

Evolutionary analysis of the whole preliminary TCGA dataset of 998 patients was carried out in PAML using the codeml program, as described in Chapter 2. This meta-analysis was carried out on a per gene basis (using whole genes as the units of analysis).

All patients were grouped together regardless of the type of mutations accumulated (mutation spectrum) and the tissue of origin (of which there were 17), before evolutionary analysis in PAML.

Chapter 5

Evolutionary sub-type analysis: stratification by tissue of origin

5.1 Introduction

The evolutionary methods used previously in Chapter 4 on the TCGA dataset were applied to the published Lawrence data, consisting of cancer-specific mutations in 4,712 patients over 21 cancer types.

The intention here was to refine the previous PAML analysis by stratifying the data by cancer type, to parse out tissue specific signals of selection, using a much larger dataset than was available for the TCGA data.

In the published work of Lawrence et al. [2014], cancer genes were also identified using this same set of cancer-specific variants. However while only the coding SNVs were used in the PAML analysis in this project, in the Lawrence et al. [2014] study both non-coding and coding mutations were used including di-, tri- and oligonucleotide variants (DNVs, TNVs and ONVs respectively) as well as INDELs.

However, Lawrence have used different methods to the ones used in this analysis, so in this sub-type analysis the aim was also to compare the genes found to be significantly

mutated in this PAML analysis to the results found by Lawrence and interpret the differences to find similarities and potentially find novel candidate genes.

5.1.1 Lawrence study

A paper was published in 2014 by Lawrence et al. [2014] in which coding and non-coding cancer-specific SNV, DNV, TNV, ONV and INDEL mutations were identified over 4,742 patients over 21 different cancer types. Described here is their criteria for defining significant genes.

5.1.1.1 MutSig software

Significantly mutated driver cancer genes were identified using the MutSig suite [Lawrence et al., 2013]. Their methods for cancer gene detection involved using three statistical tests, as is shown in Figure 5.1:

- **MutSigCV** - test for burden of protein-altering mutations, compared to a background model using "neighbouring" genes (i.e. genes that are clustered by characteristics, not by genome location).
- **MutSigCL** - mutation clustering in hotspots.
- **MutSigFN** - enrichment in functional sites, which is essentially a measure of cross-species evolutionary conservation for a site.

Each of these tests were separately applied by Lawrence et al. [2014] to each gene for each of 21 distinct tissues of origin, but the main analyses concentrated on a metric that combined these three tests into a single significance measure, the joint p-value, per gene and tissue. Of the individual MutSig tests, MutSigCV is the most directly comparable and analogous to my PAML analysis. In the MutSigCV test a background mutation model utilises both silent and non-coding mutations and combined data from multiple genes that were deemed to have similar mutation profiles. This combining of

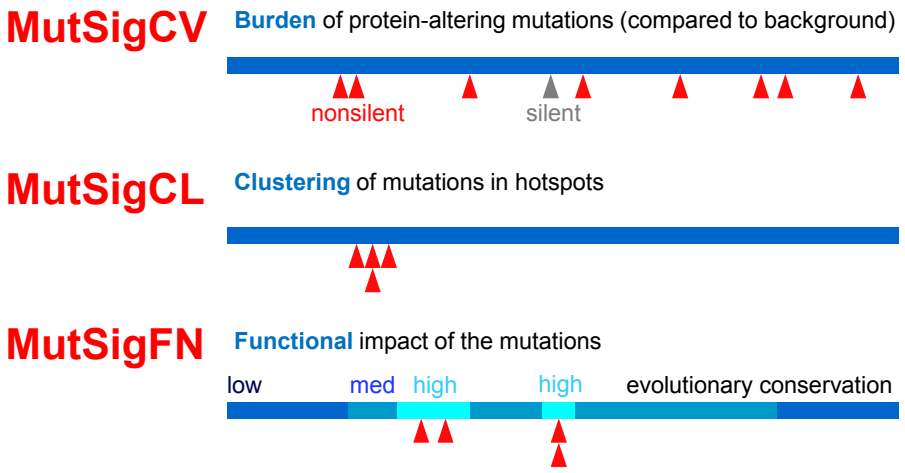


FIGURE 5.1: **Statistical tests used to detect cancer genes in MutSig.** The three statistical tests in the MutSig suite used to detect cancer genes in the Lawrence et al. [2014] study: MutSigCV, MutSigCL and MutSigFN. Schematic taken from Lawrence et al. [2014] supplementary paper.

gene background models is a potential confounding factor for the Lawrence analysis but it does serve to increase power as more "background" sites can be considered to parametrise the background model.

5.1.1.2 Data processing

Lawrence mutations were not all re-aligned to hg19 reference genome, so those originally aligned to the hg18 build were lifted over (LiftOver, see Methods) to hg19 by converting the coordinates of each mutation to build hg19.

5.1.1.3 Significance calculations: p-value and FDR

For my analysis of the Lawrence data I used log-likelihood derived p-values as a significance measure and applied a false discovery rate (FDR) correction [Benjamini and Hochberg, 1995] to account for multiple testing.

Lawrence et al. [2014] used the same q-value correction procedure, based on the joint p-values from all three MutSig tests.

In both analyses an arbitrary 10% FDR rate has been used as a threshold for significance. Genes with $q \leq 0.1$ have therefore been classified as significantly mutated.

In the Lawrence analysis, 254 genes were classed as significant and were declared to be members of the Cancer5000 list of candidate cancer genes. These genes were significant in at least one of the 22 analyses (21 distinct tissues of origin plus the combined "PanCan" set over all tumour types).

Cancer5000-S is a more stringently multiple hypothesis corrected list of genes, which has had the Benjamini-Hochberg method applied again to yield new FDR values in order to correct for the 22 analyses combined. There are 219 genes in the Cancer5000-S list.

Overall, 260 genes are present in either Cancer5000, Cancer5000-S or both.

A significance threshold of $q \leq 0.001$ was used to class genes as highly significantly mutated in Lawrence et al. [2014] (highlighted in red in subsequent plots and tables), although they have not been explicit in stating this in their paper.

5.2 Results

The Lawrence data was partitioned by tissue of origin, and of the 21 cancer types available only the 12 cancer types with at least 150 patients each were chosen for analysis in PAML, as well as melanoma which was chosen for its interesting mutation profile and high mutation rate. The PAML results for these 13 cancer types have been presented in omega plots, akin to the plot in Chapter 4 to display results from PAML analysis on the whole TCGA dataset. For each cancer type, an omega ratio and FDR was plotted for each gene.

In order to compare these results to the results in the Lawrence study, genes found to be highly significantly mutated (q<=0.001) in MutSig for that particular cancer type have been highlighted in red on the omega plots. Genes found to be significantly mutated (q<=0.1) and therefore in the Cancer5000 list of genes in the Lawrence study have been highlighted in orange in these plots.

Genes found to be significantly mutated and under strong positive selection in the PAML analysis (omega>1 and FDR<=0.1) but not significant in the Lawrence study have been highlighted in blue, and those with an omega suggestive of negative selection but a significant p-value supporting positive selection (omega<1 and FDR<=0.1) and not significant in the Lawrence et al. [2014] study have been highlighted in purple.

Ranked lists of all significantly mutated genes (omega>1 and FDR<0.1) in the PAML analysis with their omega estimate and p-value have also been tabulated for each cancer type, in descending order of significance (by ascending p-value). If these significant genes are also significant in the Lawrence et al. [2014] study then they have been highlighted according to their level of significance consistent with the omega plots: red for highly significantly mutated genes (FDR<0.001) and orange for significantly mutated genes (FDR<0.1), together with their three MutSig p-values (CL, CV, FN and combined). In addition to these significant genes in the PAML analysis, the table also includes genes that are significant in Lawrence et al. [2014] but are not classed as significant in the PAML analysis. However I have still given the omega and p-value obtained from PAML for these significant Lawrence et al. [2014] genes despite the fact the p-value is not significant in the PAML analysis (and the omega may be <1).

For clarity a horizontal blue line has been drawn on these tables to separate significant PAML genes from insignificant PAML genes. So all genes above the blue line in each table are significant in the PAML analysis (omega>1 and FDR<0.1). All genes below the line were not found to be significant in PAML analysis but are only shown in the tables because they were significant in Lawrence et al. [2014]. “NAs” in the Lawrence columns either mean that the gene has not been found to be significant in the Lawrence analysis and so p-values have not been recorded in the Lawrence study, or that p-values

were not calculated for all three MutSig tests in significant genes. "NA" in the PAML column means that PAML has not produced results for that gene. Some significant PAML genes have not been included in the tables if their Ensembl gene ID no longer exists in the most recent version of Ensembl (release 78).

Genes in the omega plots and tables have been compared based on FDR rather than p-value, which in the Lawrence dataset is based on the combined p-values from all three MutSig tests. Individual MutSig test q-values were not available, so a direct comparison was not able to be made between the MutSigCV (most directly comparable to PAML method) q-values and the PAML q-values. Therefore, since p-values were available for each of the three tests, the p-values for MutSigCV and PAML were plotted to examine the relationship between the two for each cancer type, only for genes where both a MutSigCV p-value and PAML p-value has been recorded, so this would be for the genes present in the set of the 260 most significant Lawrence genes that have also run successfully through PAML. PAML p-values are truncated at 0.5, because dchisq uses a 1-tailed test (which is suitable in this case since only positive selection has been measured), whereas MutSigCV p-values are not, however this is not observed in the plots since only the more significant p-values have been plotted. GBM showed the highest correlation so this is the only plot that has been shown here. All other cancer types did not show a strong positive correlation between the two sets of p-values.

To investigate the effectiveness of the PAML analysis, recurrent mutations (non-synonymous and synonymous) have also been counted for each cancer type and tabulated. Recurrent mutations are a partially independent method of looking for driver containing genes. The top 35 recurrent mutations (ranked in descending order of recurrence) were tabulated for each cancer type, for comparison with PAML results. Generally, genes with recurrent mutations do tend to have significant results in PAML, however several genes are found at a significant level that do not show as having a high rate of recurrent mutations, showing that a more complex model than simply counting mutations in a gene is required to detect candidate cancer genes. Also both non-synonymous and synonymous counts have been included in the recurrent counts, so this will also

confound results since synonymous mutations are considered to be selectively neutral and therefore assumed not to be driver mutations.

Synonymous mutations have been used in the background models in both this PAML analysis and in the Lawrence analysis, assumed to be passenger mutations in cancer. However it is estimated that between one in two and one in five synonymous mutations in oncogenes are under selection equating to ∼6-8% of all selected single-nucleotide changes in these genes [Supek et al., 2014], suggesting that synonymous mutations could be acting as drivers, which will confound both the PAML and Lawrence analyses.

5.2.1 Acute myeloid leukemia (LAML)

In the PAML analysis of acute myeloid leukemia, for which 194 patients were analysed, 11 genes were found to be significantly mutated with $q \leq 0.1$ and undergoing positive selection with an omega ratio >1 (top right quadrant of Figure 5.2). All 11 of these genes were also found to be highly significant in the Lawrence et al. [2014] study, with FDR$<$0.001 and have been highlighted and annotated in red in Figure 5.2. Lawrence also find additional highly significant (red) and significant genes (orange) which are not found to be significant in the PAML analysis. These are found in the bottom right quadrant of Figure 5.2, so although these genes exhibit omega ratios suggestive of positive selection they do not meet the 10% FDR threshold for positive selection.

The 11 significantly mutated genes in the PAML analysis have been tabulated in Table 5.1, showing their corresponding omega estimate and p-value. Also shown for these genes are the p-values from the MutSig analysis in the Lawrence study, for comparison. MutSig results have also been tabulated for genes found to be significantly mutated in the Lawrence study only (i.e. not significant in the PAML analysis). Overall, 26 genes were found to be significantly mutated in at least one of the two analyses.

DNMT3A is the most highly significantly mutated gene in the PAML analysis of LAML, with an omega of 475.62 and a p-value of 5.32e-72 (Table 5.1). DNMT3A is a DNA methyltransferase involved in *de novo* methylation, a process known to be involved

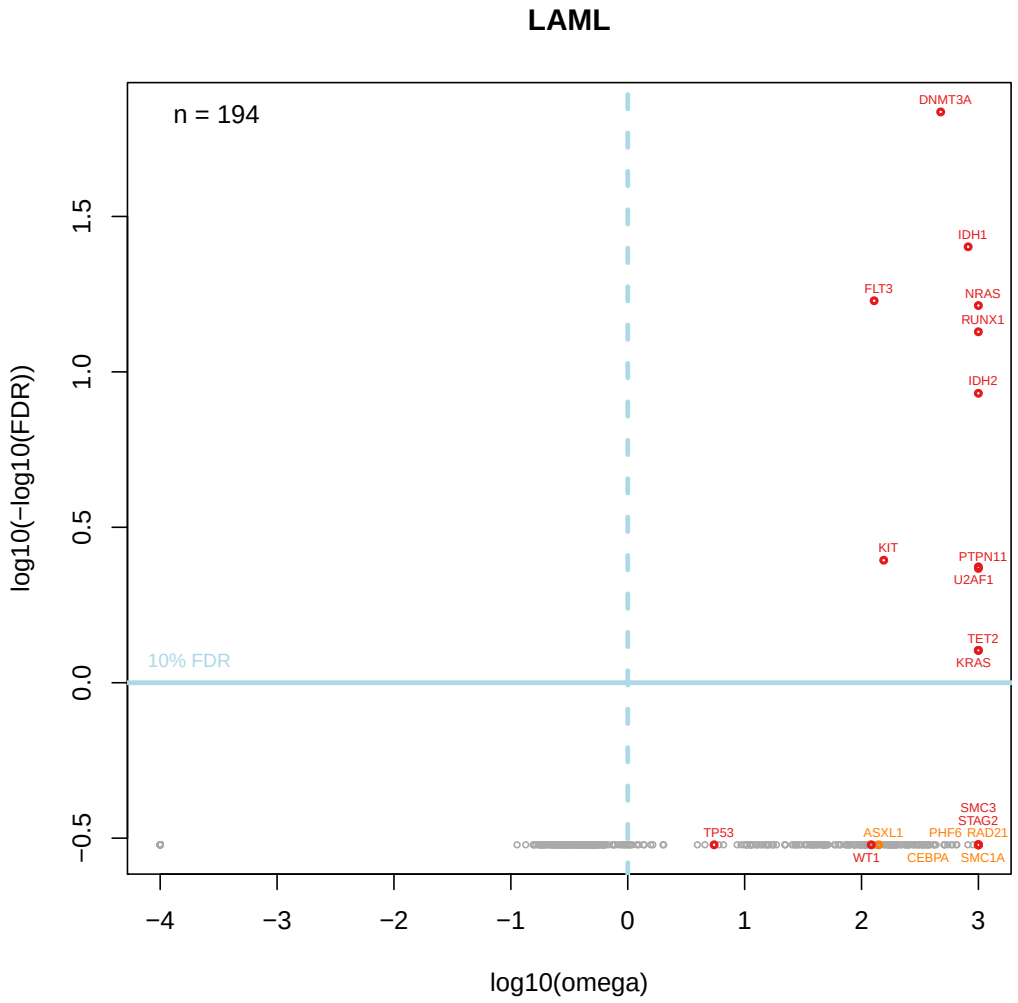


FIGURE 5.2: **Gene-based omega analysis in LAML.** Gene-based PAML results have been displayed in this omega plot for 194 LAML patients to show the omega ratios and FDR values for each gene, together with their level of significance in the Lawrence study. Genes highlighted in red and orange are those shown to be highly significantly mutated (FDR≤0.001) and significantly mutated (FDR≤0.1) respectively in the Lawrence study. *R code used to generate plot in Supplementary Appendix D.*

TABLE 5.1: **Ranked list of significant PAML genes in LAML.** The genes found to be significantly mutated in LAML patients in the PAML analysis have been sorted by p-value in ascending order (descending order of significance) in this table. Genes highlighted in red and orange are found to be highly significantly mutated and significantly mutated respectively in the Lawrence study. Genes shown below the blue horizontal line are not found to be significant in the PAML analysis but have been tabulated due to their significance in the Lawrence study. *R and Perl code used to produce list in Supplementary Appendix E.*

Gene	PAML results		Lawrence et al. [2014] p-values			
	Omega	P-value	CV	CL	FN	Combined
DNMT3A	475.62	5.32e-72	1.00E-16	9.99E-08	9.99E-08	1.11E-16
IDH1	816.36	2.27e-28	1.00E-16	4.00E-07	9.94E-01	1.11E-16
FLT3	128.20	7.24e-20	1.00E-16	9.99E-08	8.99E-07	1.11E-16
NRAS	999.00	3.73e-19	1.00E-16	9.99E-08	7.31E-02	1.11E-16
RUNX1	999.00	3.60e-16	4.02E-16	1.34E-01	1.47E-02	3.33E-16
IDH2	999.00	3.66e-11	1.00E-16	9.99E-08	3.44E-01	1.11E-16
KIT	154.77	4.84e-05	7.63E-06	4.56E-06	2.31E-01	1.05E-09
U2AF1	999.00	7.27e-05	1.70E-10	3.40E-06	9.99E-08	6.66E-16
PTPN11	999.00	8.70e-05	1.49E-11	8.70E-02	1.33E-01	1.17E-11
KRAS	999.00	1.20e-03	1.85E-13	1.14E-02	5.76E-01	6.52E-14
TET2	999.00	1.23e-03	1.00E-16	6.76E-01	1.24E-05	1.11E-16
TP53	5.48	5.98e-02	1.00E-16	1	9.20E-02	3.55E-15
SMC1A	999.00	8.65e-02	2.77E-06	1	3.30E-02	2.69E-05
SMC3	999.00	1.27e-01	6.21E-09	1	1.28E-01	9.82E-08
STAG2	999.00	1.76e-01	1.46E-08	1	8.64E-01	2.18E-07
ASXL1	140.33	2.14e-01	1.28E-06	1	8.13E-01	1.34E-05
RAD21	999.00	2.61e-01	1.56E-07	1	6.06E-01	1.96E-06
WT1	121.05	2.67e-01	5.23E-15	1.36E-03	9.95E-01	4.44E-16
PHF6	999.00	2.95e-01	1.10E-07	1	1.63E-01	1.42E-06
CEBPA	999.00	3.19e-01	7.39E-07	2.74E-01	7.28E-01	8.15E-06
PDSS2	NA	NA	2.87E-04	3.30E-02	2.71E-01	5.15E-05
EZH2	NA	NA	4.31E-04	8.27E-03	5.41E-01	3.37E-05
SFRS2	NA	NA	1.44E-05	NA	NA	1.44E-05
MXRA5	NA	NA	1.75E-03	6.73E-03	3.79E-02	8.20E-05
PAPD5	NA	NA	9.17E-05	3.40E-02	3.60E-02	2.99E-05
NPM1	NA	NA	1.00E-16	9.99E-08	9.99E-08	1.11E-16

in tumourigenesis, since aberrant promoter hypermethylation is known to be a major mechanism for silencing tumour suppressor genes in many cancers [Baylin, 2005]. DNMT3A has only recently been uncovered as an important new tumour suppressor in this cancer type [Gonzalez-Perez et al., 2013, Yang et al., 2015], although it has previously been shown to be recurrently hit with mutations in LAML patients [Ley et al., 2010]. DNMT3A has also been detected in the Lawrence analysis as highly significantly mutated in this cancer type, as well as in the combined "Pan-Cancer" dataset over all 21 cancer types [Lawrence et al., 2014].

The tumour suppressor TET2 is also found to be significant in both analyses, which could be related to the role of DNMT3A in this cancer, since TET2 acts to remove CpG methylation by converting 5-methyl-cytosine to 5-hydroxymethyl-cytosine [Yamazaki et al., 2012]. TET2 is often inactivated by loss-of-function mutations in myeloid malignancies [Solary et al., 2014]. It is possible that both DNMT3A and TET2 are targeted by driver mutations in the pathway underlying cancer progression in AML.

Compared to the genes found through evolutionary analysis, recurrent mutation analysis for LAML shown in Table 5.2 shows that 100% (11/11) of the significant PAML genes are recurrently mutated in LAML, so all significantly mutated genes in PAML can also be found using just the recurrent mutation analysis. However there are also genes containing recurrent mutations that are not significant in the PAML analysis, so using recurrent analysis alone to find significant genes would likely find many false-positives. Although not shown in Table 5.2 as it did not make the top 35 recurrent mutations, KRAS also has a recurrent mutation mutated in two patients.

Table 5.3 shows the known cancer genes that have been successfully detected in acute myeloid leukemia by both the PAML and recurrence analyses in this project, and also by the Lawrence et al. [2014] MutSig analysis.

TABLE 5.2: **Ranked list of recurrent mutations in LAML.** Ranked list of the top 35 most recurrent SNVs in the LAML subset of the Lawrence dataset, sorted by recurrence in descending order. For each unique mutation, the gene, chromosome, position, ref and alt alleles and how many patients the mutation occurs in (recurrence) have been tabulated.

Gene	Mutation				Recurrence
	Chromosome	Position	Ref	Alt	
DNMT3A	2	25457242	C	T	21
IDH2	15	90631934	C	T	16
IDH1	2	209113113	G	A	12
FLT3	13	28592642	C	A	11
DNMT3A	2	25457243	G	A	7
U2AF1	21	44524456	G	A	5
NRAS	1	115258744	C	T	5
IDH1	2	209113112	C	T	5
RUNX1	21	36231783	G	A	4
KIT	4	55599321	A	T	4
HERC6	4	89317923	A	T	4
NRAS	1	115258747	C	T	3
LPHN2	1	82434811	A	T	3
KIAA1468	18	59947868	A	T	3
IDH2	15	90631838	C	T	3
DOCK11	23	117814959	A	T	3
ABCA5	17	67280040	A	T	3
ZNF540	19	38102400	A	T	2
U2AF1	21	44524456	G	T	2
TMEM62	15	43443922	A	T	2
TET2	4	106164778	C	T	2
SPICE1	3	113225485	T	A	2
RUNX1	21	36252878	T	C	2
RDBP	6	31921657	T	A	2
RBBP4	1	33138072	G	A	2
PTPN11	12	112926884	T	C	2
POLRMT	19	630135	T	C	2
PLEKHA2	8	38826929	A	T	2
PEX5L	3	179519877	T	A	2
PDSS2	6	107566825	T	A	2
PAPD5	16	50257083	A	T	2
OR4K1	14	20403747	A	T	2
NRAS	1	115256530	G	T	2
NRAS	1	115256528	T	A	2
MXRA5	23	3229667	T	A	2

TABLE 5.3: **Cancer gene detection success in acute myeloid leukemia.** Known cancer genes that have been detected as significantly mutated in all three of the aforementioned analyses, overlapping the MutSig analysis in the Lawrence et al. [2014] study, the PAML analysis and the recurrence analysis.

Known cancer gene
FLT3
DNMT3A
IDH2
IDH1
TET2
RUNX1
NRAS
PTPN11
U2AF1
KRAS
KIT

5.2.2 Breast (BRCA)

In the PAML analysis of breast cancer, for which 888 patients were analysed, eight genes were found to be significantly mutated with q≤0.1 and undergoing positive selection with an omega ratio >1 (Figure 5.3). Of these significant genes, five were found to be highly significant in the Lawrence et al. [2014] study with a q≤0.001, and one was found to be significant (q≤0.1). The remaining two significant PAML genes, FGFR2 and CDC42BPA, were both found to be near significance in the Lawrence study, although not statistically significant. PAML analysis therefore has more power to detect these genes than MutSig, especially in the case of FGFR2 which is known to be causally implicated in cancer (Cancer Gene Census [Forbes et al., 2010]).

FGFR2 (fibroblast growth factor receptor 2) tyrosine kinase belongs to the fibroblast growth factor receptors (FGFRs) family, comprised of four kinases (FGFR1 to FGFR4) that differentially respond to 18 fibroblast growth factor (FGF) ligands [Jain and Turner, 2012]. FGFs and FGFRs play an important role in developmental signalling pathways responsible for regulating various functions such as cell proliferation, survival and migration [Rajith et al., 2013]. FGFR2 is comprised of an extracellular ligand binding region, a single-pass transmembrane region and a split kinase domain [Rajith et al., 2013]. Tyrosine kinases are enzymes which catalyse the phosphorylation of select tyrosine residues in target proteins, using ATP [Paul and Mukhopadhyay, 2004], and are commonly oncogenically activated in cancer. Constitutive oncogenic activation of tyrosine kinases in cancer cells can be blocked by selective tyrosine kinase inhibitors and thus the inhibition of these genes is considered as a promising approach for genome based therapeutics [Paul and Mukhopadhyay, 2004]. In genome-wide association studies (GWAS) a locus within the second intron in FGFR2 is consistently associated with breast cancer risk, supported by expression quantitative trait loci analysis which shows that the signalling of this gene has an important role in mediating breast cancer risk [Fletcher et al., 2013]. The predisposing allele is present in approximately 40% of western populations, although the associated increased risk is relatively small: 1.26-fold

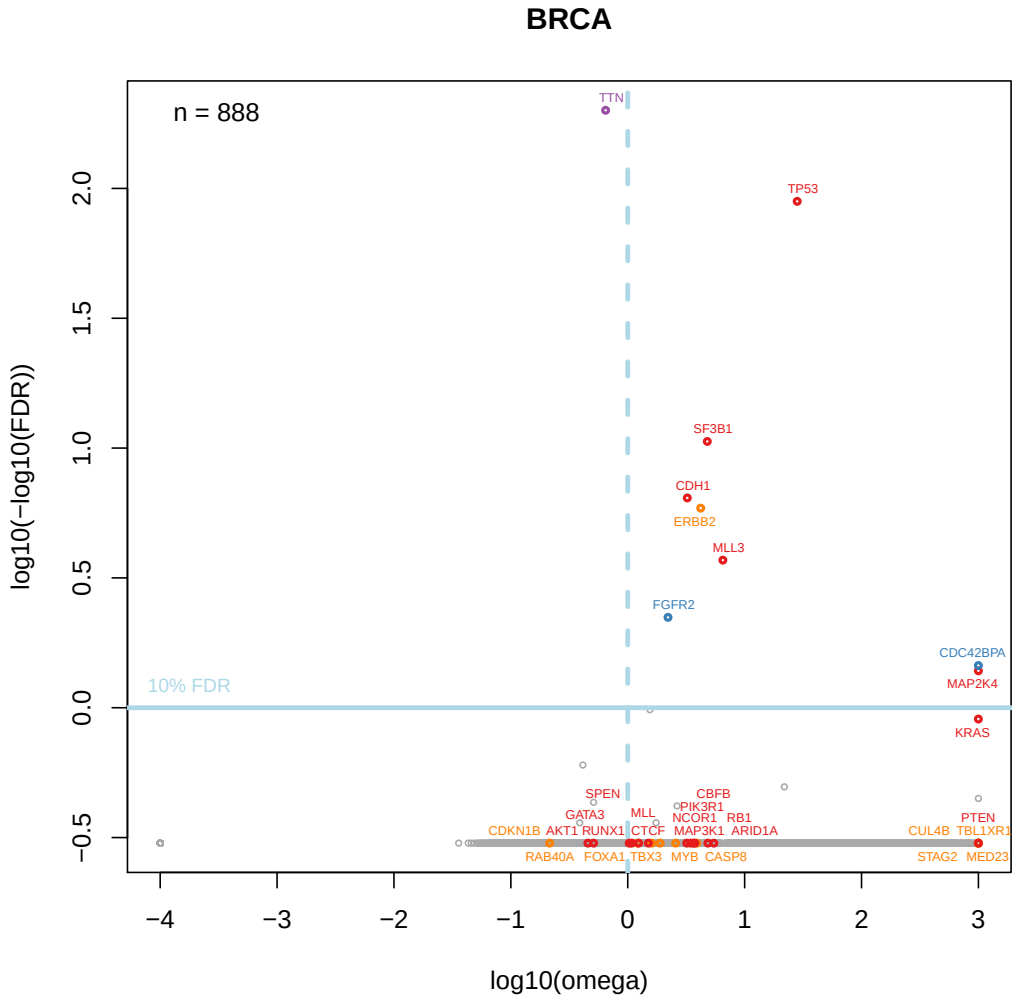


FIGURE 5.3: **Gene-based omega analysis in BRCA.** Gene-based PAML results have been displayed in this omega plot for 888 BRCA patients to show the omega ratios and FDR values for each gene, together with their level of significance in the Lawrence study. Genes highlighted in red and orange are those shown to be highly significantly mutated (FDR≤0.001) and significantly mutated (FDR≤0.1) respectively in the Lawrence study. Genes highlighted in blue are potential candidate cancer genes shown to reach significance (FDR≤0.1) in the PAML analysis, however do not reach significance in the Lawrence study. Genes highlighted in purple are examples of genes that do not reach significance in the Lawrence study, and have conflicting results in the PAML analysis (with a significant p-value supporting positive selection, but an omega value indicative of negative selection). *R code used to generate plot in Supplementary Appendix D.*

for heterozygotes and 1.63-fold for homozygotes, increasing the risk of developing oestrogen receptor (ER)-positive breast cancer, with only a minor effect on ER-negative breast cancer [Jain and Turner, 2012]. As well as germline mutations conferring breast cancer risk in FGFR2, somatic FGFR2 alterations are also known to cause constitutive activation in breast cancer Reintjes et al. [2013]. In a small subset of breast cancers the FGFR2 gene is amplified, however this is rare occurring at 1-2% of all breast cancers, (4% of aggressive triple-negative breast cancers) Jain and Turner [2012]. Breast cancers with FGFR2 gene amplification have shown high sensitivity to FGFR inhibitors and an FGFR2 targeting antibody, potentially making amplified FGFR2 a good therapeutic target. Somatic mutations have also been identified in other tumour types, with 12% of endometrial carcinomas containing somatic mutations in FGFR2 that cause oncogenic and constitutive activation. FGFR2 has also been implicated as a novel therapeutic target in endometrial carcinoma, with drugs that inhibit the kinase activity of FGFR2 shown to inhibit transformation and survival of cells [Dutt et al., 2008].

CDC42BPA belongs to the serine/threonine protein kinase family, and is not a known cancer gene. However, protein kinases are known to have an important role in cancer. This gene could therefore potentially be a good candidate for kinase inhibitors in the treatment of cancer.

TTN is the largest polypeptide encoded by the human genome (highlighted in purple in Figure 5.3) and often comes up as a false-positive in cancer studies due to the large number of mutations accumulated as a result of its large size. However, gene length has been accounted for in this analysis, by dividing by the number of non-synonymous mutations by the number of synonymous mutations to obtain a mutation rate. Therefore it may be worthwhile cautiously entertaining the possibility that this gene is a true-positive result. In the PAML results it had a FDR value of 0, which cannot be log transformed since log10(-log10(0)) gives a value of infinity. Therefore this was one of the FDR values that was changed to 1e-200 before being plotted in Figure 5.3. Although this gene has a very significant q-value supporting positive selection, it has an omega suggestive of negative selection, which has not been tested for in this analysis,

since the significance measure (FDR) is only relevant for positive selection. This could be an indication that both positive and negative selection is occurring in this gene in different parts of the gene which is confounding the omega ratio results.

Overall, 35 genes were found to be significant in at least one of the two analyses (Table 5.4). PIK3CA was found to be highly significant in the Lawrence study, however PAML was not able to produce any results for this gene (Table 5.4).

Of the significant PAML genes, 88% (7/8) were hit by recurrent mutations in BRCA, including TP53 and SF3B1 (Table 5.5). TP53 and SF3B1 were also identified in the Lawrence study as being highly significant.

Table 5.6 shows the known cancer genes that have been successfully detected in breast cancer by both the PAML and recurrence analyses in this project, and also by the Lawrence et al. [2014] MutSig analysis.

TABLE 5.4: **Ranked list of significant PAML genes in BRCA.** The genes found to be significantly mutated in BRCA patients in the PAML analysis have been sorted by p-value in ascending order (descending order of significance) in this table. Genes highlighted in red and orange are found to be highly significantly mutated and significantly mutated respectively in the Lawrence study. Genes shown below the blue horizontal line are not found to be significant in the PAML analysis but have been tabulated due to their significance in the Lawrence study. *R and Perl code used to produce list in Supplementary Appendix E.*

Gene	PAML results		Lawrence et al. [2014] p-values			
	Omega	P-value	CV	CL	FN	Combined
TP53	28.16	1.46e-93	6.01E-16	5.22E-08	5.22E-08	1.11E-16
SF3B1	4.79	8.07e-15	4.55E-05	9.89E-08	2.29E-02	1.04E-10
CDH1	3.23	1.63e-10	1.15E-15	4.00E-02	3.75E-01	3.61E-14
ERBB2	4.21	7.31e-10	5.64E-02	2.97E-07	4.75E-03	1.29E-06
MLL3 (KMT2C)	6.51	1.30e-07	1.27E-16	1.00E-01	2.92E-01	4.33E-15
FGFR2	2.21	4.50e-06	NA	NA	NA	NA
CDC42BPA	999.00	3.06e-05	NA	NA	NA	NA
MAP2K4	999.00	4.02e-05	3.52E-16	1.39E-01	2.34E-01	1.15E-14
KRAS	999.00	1.49e-04	9.98E-05	3.33E-05	1.37E-01	5.45E-08
CTCF	1.50	3.01e-03	9.13E-13	2.99E-03	2.40E-04	1.44E-15
CBFB	4.85	8.94e-03	1.00E-16	2.42E-01	3.94E-01	3.55E-15
FOXA1	1.60	1.09e-02	4.12E-06	1	2.17E-01	3.84E-05
TBX3	1.89	1.14e-02	6.82E-06	3.94E-01	9.39E-01	6.01E-05
AKT1	0.46	2.60e-02	1.00E-16	9.89E-08	1.98E-07	1.11E-16
MED23	999.00	2.62e-02	6.74E-06	1	3.15E-01	5.94E-05
RB1	3.20	3.13e-02	9.31E-09	2.99E-01	1.78E-01	1.43E-07
STAG2	999.00	3.78e-02	1.73E-05	1	7.44E-01	1.36E-04
GATA3	1.08	4.23e-02	1.00E-16	9.89E-08	3.25E-01	1.11E-16
PTEN	999.00	4.43e-02	1.00E-16	8.27E-03	1.01E-01	1.11E-16
CUL4B	999.00	9.17e-02	9.32E-06	1	6.78E-01	7.92E-05
CASP8	3.90	1.59e-01	5.79E-06	1	5.37E-01	5.19E-05
TBL1XR1	999.00	1.61e-01	6.70E-06	2.50E-02	7.94E-01	2.78E-06
HIST1H3B	3.47	1.93e-01	7.70E-05	4.85E-02	7.95E-02	1.22E-05
MAP3K1	3.65	1.96e-01	1.00E-16	1.91E-03	5.35E-01	1.11E-16
PIK3R1	3.45	2.04e-01	1.49E-10	3.16E-01	5.40E-02	2.91E-09
MYB	2.57	3.17e-01	2.02E-05	1	1.50E-02	1.56E-04
ARID1A	5.45	3.20e-01	2.66E-08	1.46E-01	1.17E-01	1.51E-08
RUNX1	1.24	4.81e-01	4.42E-16	1.08E-01	2.72E-01	1.44E-14
RAB40A	1.21	4.92e-01	2.07E-05	1	9.33E-01	1.59E-04
CDKN1B	0.21	5.00e-01	4.48E-07	1.49E-01	5.24E-01	5.17E-06
MLL (KMT2A)	1.02931	2.16e-01	2.12E-04	6.48E-04	2.70E-02	2.20E-07
SPEN	0.51	5.00e-01	2.52E-06	2.09E-03	1.46E-01	1.01E-07
NCOR1	3.75	5.00e-01	2.12E-09	1	4.74E-01	3.57E-08
PIK3CA	NA	NA	1.00E-16	9.96E-08	9.96E-08	1.11E-16

TABLE 5.5: **Ranked list of recurrent mutations in BRCA.** Ranked list of the top 35 most recurrent SNVs in the BRCA subset of the Lawrence dataset, sorted by recurrence in descending order. For each unique mutation, the gene, chromosome, position, ref and alt alleles and how many patients the mutation occurs in (recurrence) have been tabulated.

Gene	Mutation				Recurrence
	Chromosome	Position	Ref	Alt	
PIK3CA	3	178952085	A	G	119
PIK3CA	3	178936091	G	A	53
PIK3CA	3	178936082	G	A	35
AKT1	14	105246551	C	T	18
TP53	17	7578406	C	T	15
PIK3CA	3	178952085	A	T	14
PIK3CA	3	178921553	T	A	14
SF3B1	2	198266834	T	C	11
TP53	17	7578263	G	A	8
TP53	17	7578212	G	A	8
PIK3CA	3	178938934	G	A	8
TP53	17	7578271	T	C	7
TP53	17	7578190	T	C	7
TP53	17	7577539	G	A	6
TP53	17	7577121	G	A	6
PIK3CA	3	178936095	A	G	6
TP53	17	7578394	T	C	5
TP53	17	7578265	A	G	5
TP53	17	7577120	C	T	5
INF2	14	105246551	C	T	5
ZRANB3	2	136173412	A	G	4
USP6	17	5045959	T	A	4
TP53	17	7578403	C	A	4
TP53	17	7578203	C	T	4
TP53	17	7574003	G	A	4
STRN	2	37154061	A	G	4
PIK3CA	3	178936094	C	A	4
PIK3CA	3	178928079	G	A	4
PIK3CA	3	178927980	T	C	4
NUP93	16	56782199	G	A	4
MBD5	2	148909577	T	A	4
KCNJ3	2	155696027	C	A	4
FAM172A	5	93101446	C	A	4
ZNF567	19	37193020	G	A	3
ZFYVE9	1	52640752	A	C	3

TABLE 5.6: **Cancer gene detection success in breast cancer.** Known cancer genes that have been detected as significantly mutated in all three of the aforementioned analyses, overlapping the MutSig analysis in the Lawrence et al. [2014] study, the PAML analysis and the recurrence analysis.

Known cancer gene
TP53
MLL3 (KMT2C)
CDH1
MAP2K4
SF3B1
ERBB2

TABLE 5.7: **Ranked list of significant PAML genes in CLL.** The genes found to be significantly mutated in CLL patients in the PAML analysis have been sorted by p-value in ascending order (descending order of significance) in this table. Genes highlighted in red and orange are found to be highly significantly mutated and significantly mutated respectively in the Lawrence study. Genes shown below the blue horizontal line are not found to be significant in the PAML analysis but have been tabulated due to their significance in the Lawrence study. *R and Perl code used to produce list in Supplementary Appendix E.*

Gene	PAML results		Lawrence et al. [2014] p-values			
	Omega	P-value	CV	CL	FN	Combined
SF3B1	999.00	1.76e-24	2.46E-16	1.00E-07	4.75E-01	1.11E-16
MYD88	999.00	3.93e-07	5.95E-16	1.00E-07	2.40E-06	1.11E-16
TP53	999.00	1.53e-06	1.00E-16	2.81E-03	1.00E-07	1.11E-16
XPO1	1.02	4.38e-03	5.37E-06	3.90E-04	1.90E-01	3.54E-08
HIST1H1E	999.00	5.03e-02	1.36E-06	1	7.73E-03	6.26E-07
RPS15	999.00	3.09e-01	1.67E-07	1	5.07E-01	2.09E-06
RPS2	21.24	4.08e-01	1.24E-04	1.50E-02	8.30E-01	1.89E-05

5.2.3 Chronic lymphocytic leukemia (CLL)

In the PAML analysis of chronic lymphocytic leukemia, for which 159 patients were analysed, three genes were found to be significantly mutated with q≤0.1 and undergoing positive selection with an omega ratio >1 (Figure 5.4). All three of these genes were also found to be highly significantly mutated in the Lawrence et al. [2014] study.

Overall, just seven genes were found to be significantly mutated in at least one of the analyses (Table 5.7).

In Table 5.8 100% (3/3) of the genes found to be significantly mutated in the PAML analysis were also found to contain recurrent mutations in CLL.

Table 5.9 shows the known cancer genes that have been successfully detected in chronic lymphocytic leukemia by both the PAML and recurrence analyses in this project, and also by the Lawrence et al. [2014] MutSig analysis.

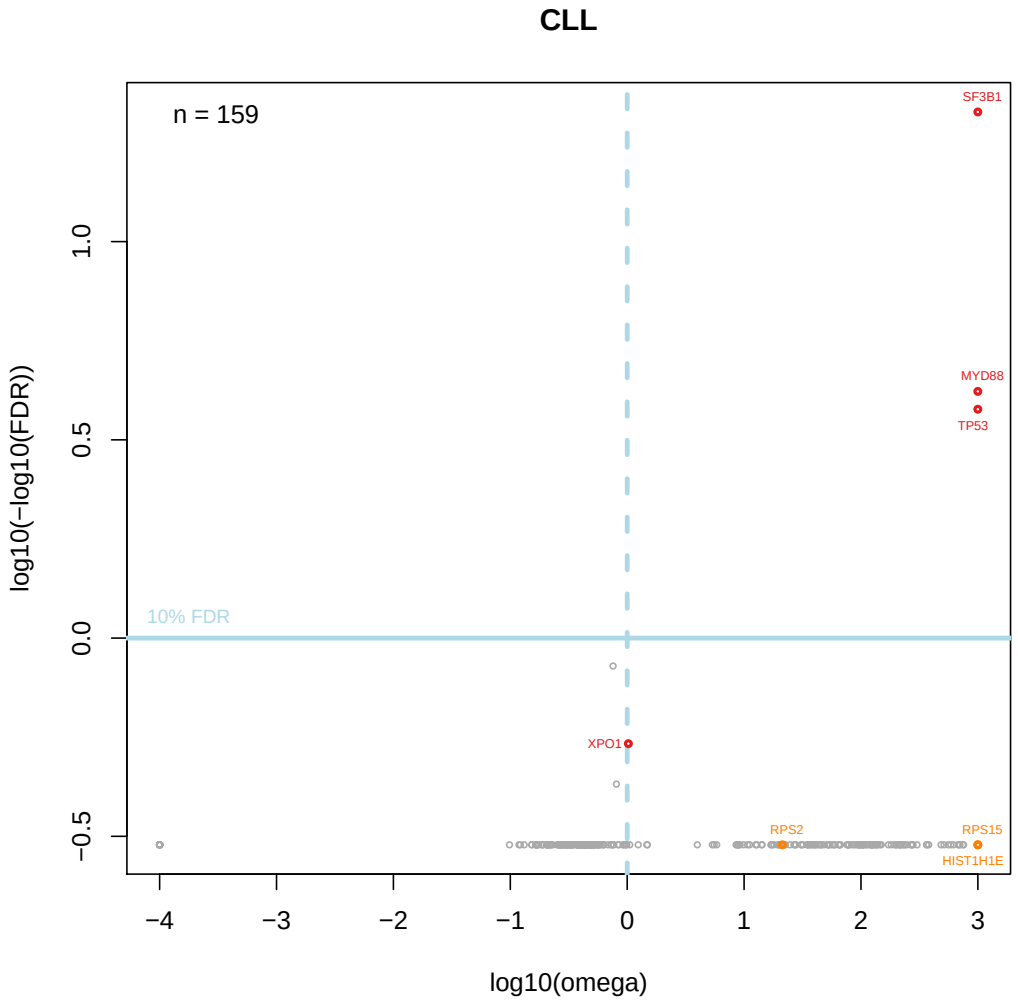


FIGURE 5.4: **Gene-based omega analysis in CLL.** Gene-based PAML results have been displayed in this omega plot for 159 CLL patients to show the omega ratios and FDR values for each gene, together with their level of significance in the Lawrence study. Genes highlighted in red and orange are those shown to be highly significantly mutated (FDR≤0.001) and significantly mutated (FDR≤0.1) respectively in the Lawrence study. *R code used to generate plot in Supplementary Appendix D.*

TABLE 5.8: **Ranked list of recurrent mutations in CLL.** Ranked list of the 13 recurrent SNVs in the CLL subset of the Lawrence dataset, sorted by recurrence in descending order. For each unique mutation, the gene, chromosome, position, ref and alt alleles and how many patients the mutation occurs in (recurrence) have been tabulated.

Gene	Mutation				Recurrence
	Chromosome	Position	Ref	Alt	
SF3B1	2	198266834	T	C	9
MYD88	3	38182641	T	C	9
XPO1	2	61719472	C	T	3
SF3B1	2	198266611	C	T	3
VWF	12	6125705	C	A	2
TP53	17	7577556	C	A	2
TP53	17	7577141	C	A	2
TP53	17	7577120	C	T	2
RPS2	16	2012609	T	C	2
NRAS	1	115256529	T	C	2
MYD88	3	38182259	T	C	2
GNB1	1	1737942	A	G	2
EGR2	10	64573248	G	T	2

TABLE 5.9: **Cancer gene detection success in chronic lymphocytic leukemia.** Known cancer genes that have been detected as significantly mutated in all three of the aforementioned analyses, overlapping the MutSig analysis in the Lawrence et al. [2014] study, the PAML analysis and the recurrence analysis.

Known cancer gene
SF3B1
TP53
MYD88

5.2.4 Colorectal (CRC)

In the PAML analysis of colorectal cancer, for which 233 patients were analysed, 71 genes were found to be significantly mutated with q≤0.1 and undergoing positive selection with an omega ratio >1 (Figure 5.5). The most significant of these genes is APC, mutations in which are known to be the most common inactivation in this type of cancer [Fodde, 2002]. This gene is also shown to be highly significantly mutated in Lawrence. This result shows that both methods have the power to successfully detect this cancer gene.

Overall 80 genes were found to be significantly mutated in at least one of the two analyses (Table 5.10).

Of all the genes found to be highly significantly mutated in the Lawrence study, all but BRAF were also found to be significant in the PAML analysis. A possible explanation for this could be the presence of clustering of missense mutations in this gene observed in Figure 5.6. MutSig uses a test called MutSigCL that specifically measures and accounts for clustering in hotspots within genes. This has not been accounted for in the PAML method, which could explain why BRAF has not come up as a significant gene in PAML analysis.

A candidate cancer gene that has been shown to be significantly mutated in PAML analysis but not in the Lawrence analysis is DNMT1. This gene is involved in DNA methylation (the major form of epigenetic information in mammalian cells), which is a methyl transfer reaction performed by trans-acting enzymes known as DNA methlytrasferases such as DNMT1 involving the covalent addition of a methly group to the 5'-position of cytosince predominantly within the CpG dinucleotide [Robertson, 2001]. Two distinct methyl transfer activities can be distinguished, based on the methylation status of the substrate: maintenance DNA methyltrasferase activity refers to the conversion of hemimethlyated substrates to a fully methylated state; whereas *de novo* methyltransferase activity refers to the addition of new methyl groups at sites previously unmethlyated. All DNA methyltransferases are capable of performing both

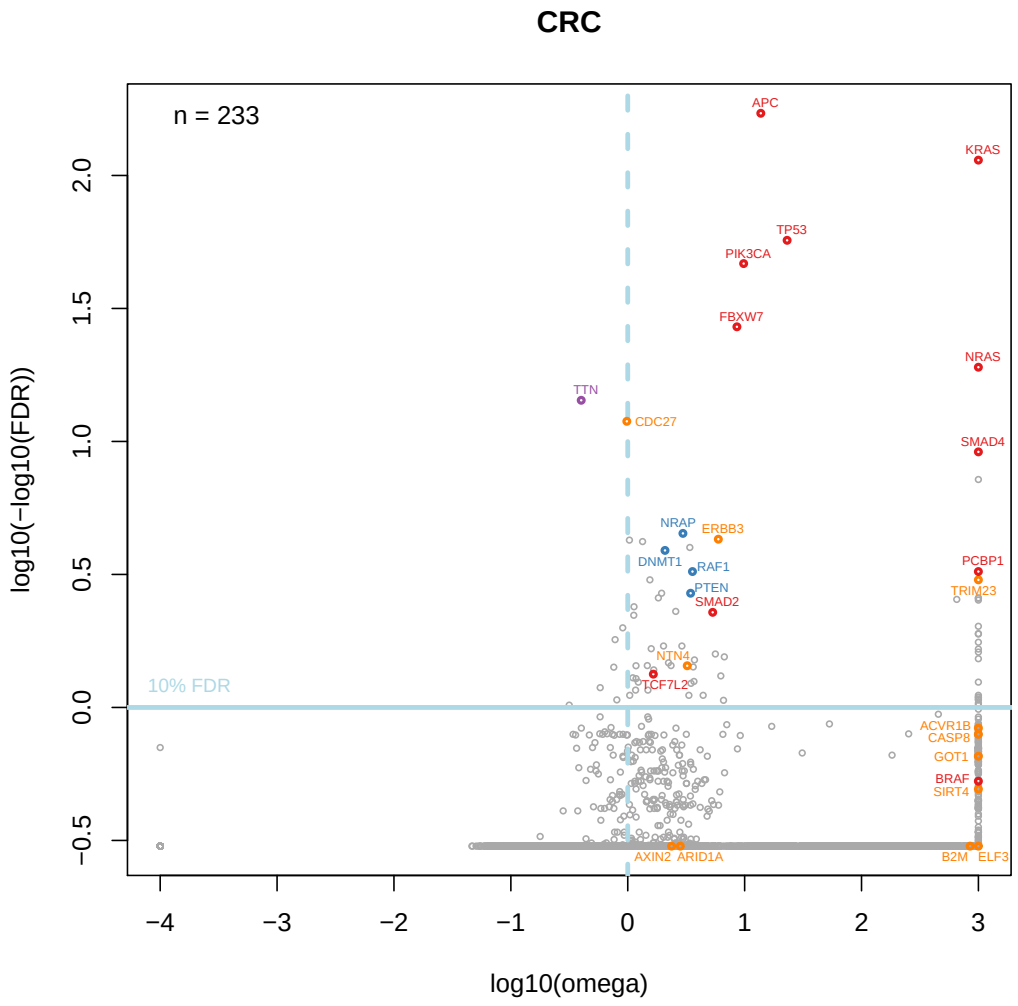


FIGURE 5.5: **Gene-based omega analysis in CRC.** Gene-based PAML results have been displayed in this omega plot for 233 CRC patients to show the omega ratios and FDR values for each gene, together with their level of significance in the Lawrence study. Genes highlighted in red and orange are those shown to be highly significantly mutated (FDR≤0.001) and significantly mutated (FDR≤0.1) respectively in the Lawrence study. Genes highlighted in blue are potential candidate cancer genes shown to reach significance (FDR≤0.1) in the PAML analysis, however do not reach significance in the Lawrence study. Genes highlighted in purple are examples of genes that do not reach significance in the Lawrence study, and have conflicting results in the PAML analysis (with a significant p-value supporting positive selection, but an omega value indicative of negative selection). *R code used to generate plot in Supplementary Appendix D.*

TABLE 5.10: **Ranked list of significant PAML genes in CRC.** The genes found to be significantly mutated in CRC patients in the PAML analysis have been sorted by p-value in ascending order (descending order of significance) in this table. Genes highlighted in red and orange are found to be highly significantly mutated and significantly mutated respectively in the Lawrence study. Table has been truncated to n=35 rows from a total of 80 significant genes for CRC. *R and Perl code used to produce list in Supplementary Appendix E. Full version of table can be found in Supplementary Appendix F.*

	PAML results		Lawrence et al. [2014] p-values			
Gene	Omega	P-value	CV	CL	FN	Combined
APC	13.75	3.65e-176	1.26E-15	9.28E-08	9.98E-01	1.11E-16
KRAS	999.00	1.00e-118	6.79E-04	9.00E-08	4.23E-06	1.25E-09
TP53	23.09	2.20e-61	4.53E-15	2.34E-06	9.00E-08	1.11E-16
PIK3CA	9.81	7.38e-51	5.94E-03	9.00E-08	1.15E-04	9.76E-09
FBXW7	8.59	4.26e-31	1.00E-16	7.74E-04	1.89E-01	1.11E-16
NRAS	999.00	4.75e-23	3.87E-16	9.00E-08	2.76E-03	1.11E-16
CDC27	0.98	8.16e-16	3.44E-01	2.46E-06	9.98E-01	5.00E-05
SMAD4	999.00	5.36e-13	1.27E-15	7.26E-03	1.10E-03	1.11E-16
CTC-554D6.1	999.00	5.24e-11	NA	NA	NA	NA
NRAP	2.97	2.77e-08	NA	NA	NA	NA
ERBB3	5.95	5.03e-08	5.96E-04	7.07E-04	4.47E-01	5.80E-06
PPP2R1A	1.03	5.90e-08	NA	NA	NA	NA
ERBB2	1.34	7.17e-08	NA	NA	NA	NA
ABCA8	3.40	1.25e-07	NA	NA	NA	NA
DNMT1	2.08	1.68e-07	NA	NA	NA	NA
RAF1	3.58	8.34e-07	NA	NA	NA	NA
PCBP1	999.00	8.44e-07	4.71E-03	1.50E-05	1.35E-03	2.76E-07
HLCS	1.55	1.55e-06	NA	NA	NA	NA
TRIM23	999.00	1.57e-06	7.85E-01	1.47E-05	1.03E-01	4.47E-05
PTEN	3.45	3.65e-06	NA	NA	NA	NA
MYO3A	1.95	3.70e-06	NA	NA	NA	NA
DNAH5	1.83	5.13e-06	NA	NA	NA	NA
CUL2	999.00	5.15e-06	NA	NA	NA	NA
MORC2	999.00	5.48e-06	NA	NA	NA	NA
ACPP	652.61	6.03e-06	NA	NA	NA	NA
SFI1	999.00	6.49e-06	NA	NA	NA	NA
NEB	1.13	9.33e-06	NA	NA	NA	NA
DKK2	2.57	1.20e-05	NA	NA	NA	NA
SMAD2	5.33	1.30e-05	1.83E-11	1.39E-03	1.98E-02	7.55E-14
CLEC18C	1.13	1.53e-05	NA	NA	NA	NA
NEDD9	999.00	2.51e-05	NA	NA	NA	NA
RBBP7	999.00	3.55e-05	NA	NA	NA	NA
GPHN	999.00	3.71e-05	NA	NA	NA	NA
ZNHIT6	999.00	5.28e-05	NA	NA	NA	NA
PTPDC1	2.91	6.31e-05	NA	NA	NA	NA

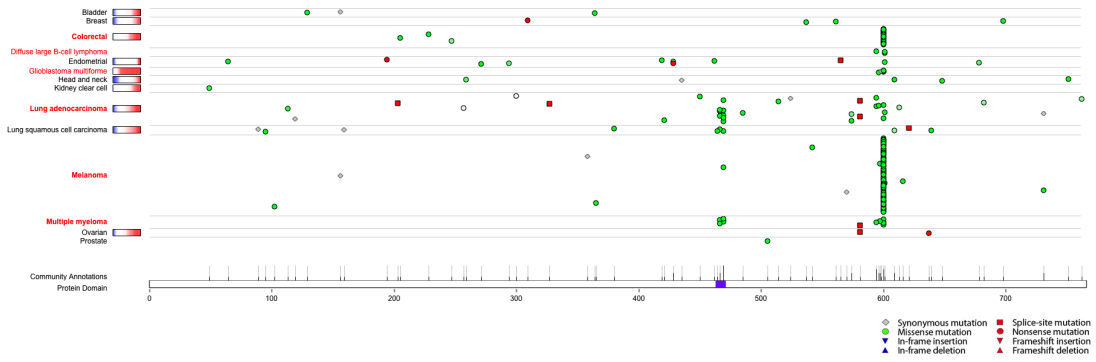


FIGURE 5.6: **Missense mutation clustering in BRAF.** Mutation pattern in BRAF with mutation key. This plot shows the clustering of missense mutations that is occurring in the BRAF gene at a specific position over all cancer types, taken from URL: http://cancergenome.broadinstitute.org/. This could explain why MutSig is able to detect this gene as significant in Lawrence et al. [2014], and why PAML is not.

reactions, however the predominant mammalian DNA methyltransferase DNMT1 is unusual in that its relative *de novo* activity is 1-2 orders of magnitude lower than its maintenance activity, and is known as the 'maintenance' methyltransferase since it is believed to be the primary enzyme responsible for copying methylation patterns after DNA replication [Robertson, 2001]. Other DNA methyltransferases such as DNMT3A and DNMT3B have approximately equal ratios of *de novo* methyltransferase activity:maintenance DNA methyltransferase activity. DNMT1 was the first methyltransferase to be identified [Bestor et al., 1988], and is the most abundant methlytransferase in somatic cells [Robertson, 2001].

Genomic methylation patterns are frequently altered in tumour cells, and when hypermethylation events occur within the promoter of a tumour suppressor gene expression of the associated gene can be silenced and provide the cell with a selective growth advantage [Robertson, 2001], hence why this gene is of great interest.

DNMT1 has long been suspected as a candidate cancer gene, particularly in this tumour type, but has never been confirmed as one. For example, Eads et al. [1999] investigated the molecular basis of aberrant *de novo* hypermethylation of CpG islands observed in a a subset of human colorectal tumours, hypothesising that one potential mechanism was through the up-regulation of DNA (cytosine-5')-methyltransferases such as DNMT1.

However, it was concluded that the deregulation of DNA methyltransferase gene expression did not play a role in establishing tumour-specific abnormal DNA methylation patterns in human colorectal cancer.

Three sites are recurrently mutated in DNMT1 in the Lawrence colorectal subset: E432K, A544P and E1531Q. Almost all of the changes are non-synonymous point mutations, and so not necessarily knock-outs, since mutations that knock out function of a gene are typically more likely to be in the form of frameshift indels, nonsense point mutations that introduce a stop codon or large deletions. Therefore these mutations could be either activating or inactivating the function of the protein encoded by the gene. Since abnormal methlyation of CpG islands associated with tumour suppressor genes can lead to transcriptional silencing [Eads et al., 1999], inactivating the gene through epigenetic means, it can be hypothesised that these putative driver mutations in DNMT1 are activating mutations that re up-regulating methyltransferase activity to cause hypermethylation of a tumour suppressor gene substrate. The mutations localise to two regions of the protein: N-terminal interaction domains and C-terminal catalytic domains.

The proto-oncogene serine/threonine-protein kinase, RAF1, is also significant in the PAML analysis but not in the Lawrence study. However it has previously been implicated in colorectal cancer, known to be important in the RAS/MAPK1 signaling pathway, and is involved in the control of cell proliferation [Slattery et al., 2012]. This is a cancer gene that Lawrence has failed to detect, but that PAML has successfully identified.

Of all the genes highly significantly mutated in Lawrence, PCBP1 is the only gene not known to be implicated in cancer. This gene was also detected by the PAML analysis, so in this case both methods have detected a novel candidate cancer gene.

91% (64/70) of the significant genes in the PAML analysis of CRC were hit by recurrent mutations. Mutations in the top most significantly mutated genes in PAML, APC and KRAS, are amongst the top 35 most recurrent mutations in Table 5.11, with one

particular mutation in APC occurring in 19 patients and a KRAS mutation occurring 31 times across all patients.

Table 5.12 shows the known cancer genes that have been successfully detected in colorectal cancer by both the PAML and recurrence analyses in this project, and also by the Lawrence et al. [2014] MutSig analysis.

TABLE 5.11: **Ranked list of recurrent mutations in CRC.** Ranked list of the top 35 most recurrent SNVs in the CRC subset of the Lawrence dataset, sorted by recurrence in descending order. For each unique mutation, the gene, chromosome, position, ref and alt alleles and how many patients the mutation occurs in (recurrence) have been tabulated.

Gene	Mutation				Recurrence
	Chromosome	Position	Ref	Alt	
KRAS	12	25398284	C	T	31
KRAS	12	25398284	C	A	23
BRAF	7	140453136	A	T	20
APC	5	112175639	C	T	19
TP53	17	7578406	C	T	15
KRAS	12	25398281	C	T	15
APC	5	112173917	C	T	12
PIK3CA	3	178936091	G	A	9
TP53	17	7577539	G	A	8
PIK3CA	3	178952085	A	G	8
KRAS	12	25378562	C	T	8
FBXW7	4	153249384	C	T	8
TP53	17	7578212	G	A	7
TP53	17	7577120	C	T	7
SMAD4	18	48591919	G	A	7
CDC27	17	45216162	A	C	7
APC	5	112173704	C	T	7
APC	5	112128143	C	T	7
NRAS	1	115256530	G	T	6
KRAS	12	25398285	C	A	6
APC	5	112175423	C	T	6
APC	5	112174631	C	T	6
APC	5	112164616	C	T	6
APC	5	112116592	C	T	5
TP53	17	7578263	G	A	4
TP53	17	7577548	C	T	4
TP53	17	7577121	G	A	4
TP53	17	7577022	G	A	4
PIK3CA	3	178936082	G	A	4
PIK3CA	3	178916876	G	A	4
FBXW7	4	153249385	G	A	4
ERBB3	12	56478854	G	A	4
ERBB2	17	37881332	G	A	4
EMR3	19	14772866	G	A	4
APC	5	112175390	C	T	4

TABLE 5.12: **Cancer gene detection success in colorectal cancer.** Known cancer genes that have been detected as significantly mutated in all three of the aforementioned analyses, overlapping the MutSig analysis in the Lawrence et al. [2014] study, the PAML analysis and the recurrence analysis.

Known cancer gene
APC
TP53
KRAS
PIK3CA
FBXW7
SMAD4
TCF7L2
NRAS
SMAD2
ERBB3

5.2.5 Endometrial (UCEC)

In the PAML analysis of endometrial cancer, for which 247 patients were analysed, 666 genes were found to be significantly mutated with q≤0.1 and undergoing positive selection with an omega ratio >1 (Figure 5.7).

UCEC is known for its high mutation rate across all genes, which explains why there is such a high rate of significant gene detection in the PAML analysis. The Lawrence analysis has identified far fewer significant genes in comparison, which suggests that PAML is less conservative. However, all 16 of the genes found to be highly significantly mutated in the Lawrence study have also been found to be significant in the PAML analysis, which shows good synergy between the two methods.

Of the 666 significant PAML genes, DNAH5 is one that has not been detected by the MutSig analysis. This gene encodes outer dynein arm components, and mutations in this gene are a common cause of primary ciliary dyskinesia with outer dynein arm defects and is frequently mutated in patients with myeloma [Hornef et al., 2006]. However, DNAH5 is not a known cancer gene in UCEC or in any other cancer type, and [Lawrence et al., 2014] did not detect this gene in any of their 22 analyses. Recently, however, this gene has been shown to potentially play an important role in colorectal cancer, shown to be downregulated in both colon and rectal cancers compared to normal control tissues [Xiao et al., 2015]. DNMT1 could therefore represent a novel candidate cancer gene in endometrial cancer, and has been considered as a potential candidate biomarker for diagnosis, prognosis and as a therapeutic target for this particular malignancy.

To illustrate the power of PAML to detect known cancer genes, MYC associated factor X (MAX) was detected as significantly significant by the PAML analysis, but was not significant in the MutSig analysis. This gene is known to be associated with the following cancers: pheochromocytoma; endometrioid carcinoma; and colon carcinoma [Futreal et al., 2004]. Tumour-specific inactivation of this gene was discovered in small cell lung cancers, and in MAX-deficient lung cancers the inactivation of the chromatin remodeler BRG1 revealed a preferential toxicity, suggesting a novel therapeutic target

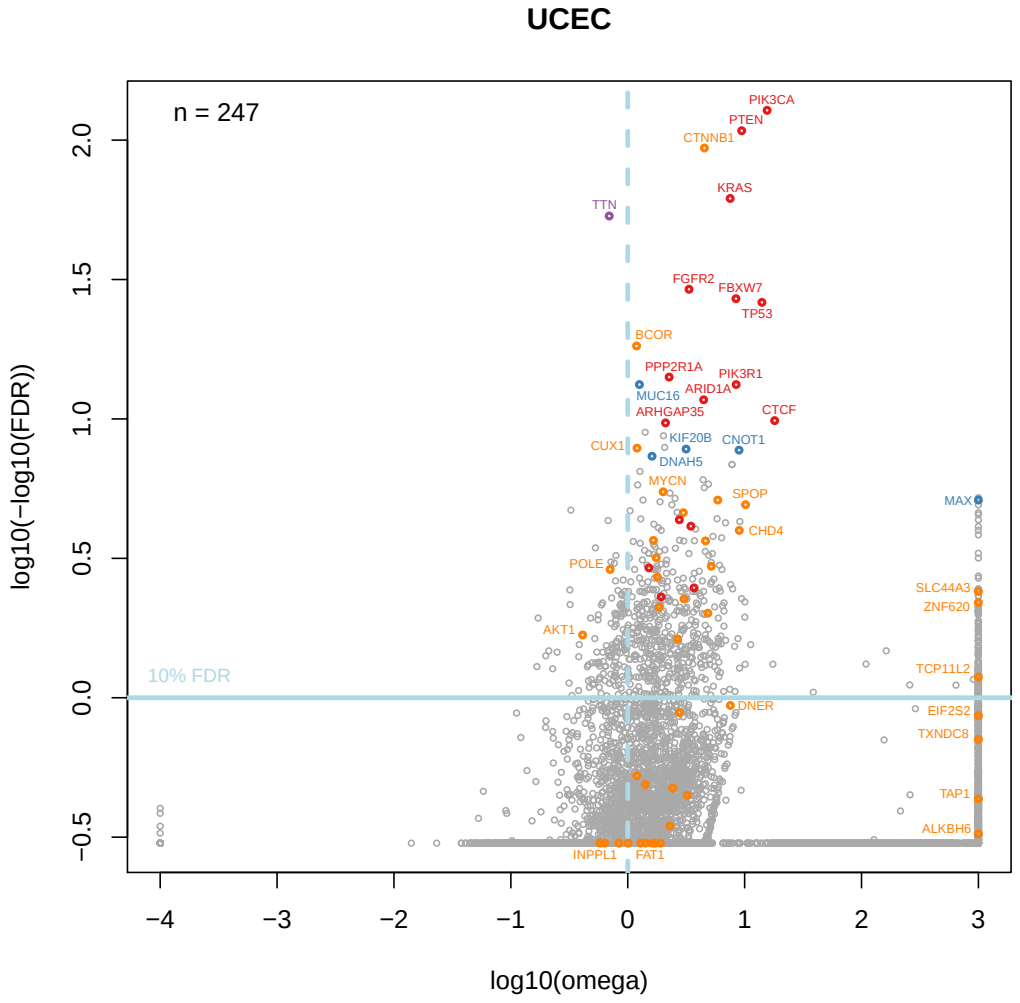


FIGURE 5.7: **Gene-based omega analysis in UCEC.** Gene-based PAML results have been displayed in this omega plot for 247 UCEC patients to show the omega ratios and FDR values for each gene, together with their level of significance in the Lawrence study. Genes highlighted in red and orange are those shown to be highly significantly mutated (FDR≤0.001) and significantly mutated (FDR≤0.1) respectively in the Lawrence study. Genes highlighted in blue are potential candidate cancer genes shown to reach significance (FDR≤0.1) in the PAML analysis, however do not reach significance in the Lawrence study. Genes highlighted in purple are examples of genes that do not reach significance in the Lawrence study, and have conflicting results in the PAML analysis (with a significant p-value supporting positive selection, but an omega value indicative of negative selection). *R code used to generate plot in Supplementary Appendix D.*

for the treatment of patients with MAX-deficient tumours [Romero et al., 2014]. If this gene functions through a similar pathway in endometrial cancer then this could suggest a putative role of MAX as a candidate cancer gene in UCEC.

Overall, 687 genes were found to be significant across both studies (Table 5.13).

87% (577/666) of the significant genes found in UCEC through PAML analysis are hit by recurrent mutations. The top five most significantly mutated genes from PAML are PIK3CA, PTEN, CTNNB1, KRAS and FGFR2. These genes are also found to be hit by some of the most recurrent mutations in this cancer type (Table 5.14).

Table 5.15 shows the known cancer genes that have been successfully detected in endometrial cancer by both the PAML and recurrence analyses in this project, and also by the Lawrence et al. [2014] MutSig analysis.

TABLE 5.13: **Ranked list of significant PAML genes in UCEC.** The genes found to be significantly mutated in UCEC patients in the PAML analysis have been sorted by p-value in ascending order (descending order of significance) in this table. Genes highlighted in red and orange are found to be highly significantly mutated and significantly mutated respectively in the Lawrence study. Table has been truncated to n=35 rows from a total of 687 significant genes for UCEC. *R and Perl code used to produce list in Supplementary Appendix E. Full version of table can be found in Supplementary Appendix F.*

	PAML results		Lawrence et al. [2014] p-values			
Gene	Omega	P-value	CV	CL	FN	Combined
PIK3CA	15.57	7.80e-133	1.00E-16	9.41E-08	5.64E-07	1.11E-16
PTEN	9.43	9.90e-113	1.85E-16	6.83E-08	8.28E-01	1.11E-16
CTNNB1	4.53	3.98e-98	6.34E-02	9.00E-08	9.00E-08	1.18E-06
KRAS	7.52	3.99e-66	1.00E-16	5.02E-08	1.12E-03	1.11E-16
FGFR2	3.34	2.39e-33	1.12E-03	5.01E-08	4.85E-02	1.15E-09
FBXW7	8.43	4.24e-31	7.14E-16	6.25E-03	1.27E-02	1.11E-16
TP53	14.06	3.14e-30	2.55E-15	1.94E-07	6.48E-08	1.11E-16
BCOR	1.19	2.87e-22	2.97E-01	5.99E-05	1.19E-04	1.29E-06
PPP2R1A	2.26	4.47e-18	8.01E-03	4.32E-07	1.01E-01	5.01E-08
MUC16	1.26	3.64e-17	NA	NA	NA	NA
PIK3R1	8.46	3.79e-17	8.18E-16	9.57E-08	4.46E-01	1.11E-16
ARID1A	4.46	1.53e-15	1.00E-16	9.52E-01	6.99E-01	3.55E-15
CTCF	18.07	1.16e-13	1.00E-16	6.00E-03	3.94E-01	2.22E-16
ARHGAP35	2.10	1.89e-13	2.05E-13	6.36E-01	2.06E-01	5.35E-12
KIAA2026	1.41	1.06e-12	NA	NA	NA	NA
NEB	2.02	2.01e-12	NA	NA	NA	NA
ASCC3	2.07	1.35e-11	NA	NA	NA	NA
CUX1	1.20	1.59e-11	3.26E-07	6.90E-02	9.95E-01	3.86E-06
KIF20B	3.16	1.92e-11	NA	NA	NA	NA
CNOT1	8.96	2.37e-11	NA	NA	NA	NA
DNAH5	1.61	5.94e-11	NA	NA	NA	NA
ZNF709	7.80	1.98e-10	NA	NA	NA	NA
CCDC132	1.27	4.96e-10	NA	NA	NA	NA
ZNF781	4.41	1.39e-09	NA	NA	NA	NA
TXNL1	4.90	2.38e-09	NA	NA	NA	NA
SMC4	1.22	2.69e-09	NA	NA	NA	NA
TAF1	4.52	3.77e-09	NA	NA	NA	NA
MYCN	2.01	6.03e-09	1.50E-01	2.53E-05	6.41E-02	1.97E-04
MDN1	2.30	7.14e-09	NA	NA	NA	NA
SFRP4	999.00	1.17e-08	NA	NA	NA	NA
INTS7	2.54	1.28e-08	NA	NA	NA	NA
ZNF43	1.35	1.58e-08	NA	NA	NA	NA
MAX	999.00	1.65e-08	NA	NA	NA	NA
RSBN1L	5.89	1.66e-08	1.21E-02	3.28E-04	8.83E-01	1.03E-04
ZNF180	1.89	2.01e-08	NA	NA	NA	NA

TABLE 5.14: **Ranked list of recurrent mutations in UCEC.** Ranked list of the top 35 most recurrent SNVs in the UCEC subset of the Lawrence dataset, sorted by recurrence in descending order. For each unique mutation, the gene, chromosome, position, ref and alt alleles and how many patients the mutation occurs in (recurrence) have been tabulated.

Gene	Mutation				Recurrence
	Chromosome	Position	Ref	Alt	
PTEN	10	89692904	C	G	25
PTEN	10	89692905	G	A	17
KRAS	12	25398284	C	T	16
PTEN	10	89717672	C	T	14
FGFR2	10	123279677	G	C	14
BCOR	23	39921444	T	C	13
PIK3CA	3	178952085	A	G	12
PIK3CA	3	178916876	G	A	12
KRAS	12	25398284	C	A	12
PIK3CA	3	178936091	G	A	10
PPP2R1A	19	52715971	C	G	9
PIK3CA	3	178936082	G	A	9
ARID1A	1	27106354	C	T	9
POLE	12	133253184	G	C	8
PIK3R1	5	67588951	C	T	8
CTNNB1	3	41266113	C	T	8
CTNNB1	3	41266113	C	G	8
PTEN	10	89692904	C	T	7
KRAS	12	25398281	C	T	7
PIK3CA	3	178952085	A	T	6
PIK3CA	3	178917478	G	A	6
KIAA2026	9	5968511	G	A	6
CTNNB1	3	41266101	C	T	6
CTNNB1	3	41266097	G	A	6
CTCF	16	67655479	C	T	6
ARHGAP35	19	47424921	C	T	6
TPP2	13	103292684	C	T	5
TP53	17	7577539	G	A	5
SMC4	3	160151484	C	T	5
SLCO1B3	12	21011428	G	A	5
POLE	12	133250289	C	A	5
PIK3CA	3	178916891	G	A	5
MYO10	5	16694613	C	T	5
MYF6	12	81101837	C	T	5
KRAS	12	25398284	C	G	5

TABLE 5.15: **Cancer gene detection success in endometrial cancer.** Known cancer genes that have been detected as significantly mutated in all three of the aforementioned analyses, overlapping the MutSig analysis in the Lawrence et al. [2014] study, the PAML analysis and the recurrence analysis.

Known cancer gene
PTEN
PIK3CA
PIK3R1
ARID1A
TP53
KRAS
CTCF
ZFHX3
FBXW7
FGFR2
PPP2R1A
CCND1
ERBB3
ING1
CTNNB1
BCOR
CHD4
CUX1
SPOP
SOX17
NRAS
MYCN

5.2.6 Glioblastoma multiforme (GBM)

In the PAML analysis of glioblastoma multiforme, for which 291 patients were analysed, 20 genes were found to be significantly mutated with q≤0.1 and undergoing positive selection with an omega ratio >1 (Figure 5.8).

A potential candidate cancer gene called HERC1 has been identified by the PAML analysis, undetected by MutSig. The HERC1 protein belongs to the HERC protein family of ubiquitin ligases characterized by the presence of an HECT domain and one or more RCC1-like domains [Nguyen et al., 2015]. This family consists of two subgroups according to their sizes and domain architecture, with HERC1 as one of the two giant HERC proteins containing other functional domains such as two RLDs, a C-terminal HECT, a WD40, a SPRY (spl A and RyR) domain and several other minor motifs [Nguyen et al., 2015]. A homozygous nonsense variant in HERC1 was identified through exome sequencing as the causal variant of a novel autosomal-recessive neurological condition characterized by megalencephaly, thick corpus callosum and severe intellectual disability. HERC1 is also found to interact with an upstream negative regulator of the mTOR pathway, which has been shown to have a role in the regulation of brain development [Nguyen et al., 2015]. The role of this gene in a pathway of the brain and with a mutation known to cause disease specifically in this tissue supports this gene as a candidate for glioblastoma mutliforme.

Overall, 32 genes were found to be significant in at least one of the two analyses (Table 5.16).

Figure 5.9 shows the relationship between the PAML and MutSig p-values for the most significant genes. A strong positive correlation is observed between the two p-values (Pearson's cor = 0.61, p-value = 1.76e-11).

Again significant PAML genes are shown to be genes containing recurrent mutations for this cancer type, with 95% (19/20) of the significant genes containing recurrent mutations. For example EGFR, TP53, PTEN and PIK3CA are all significantly mutated in

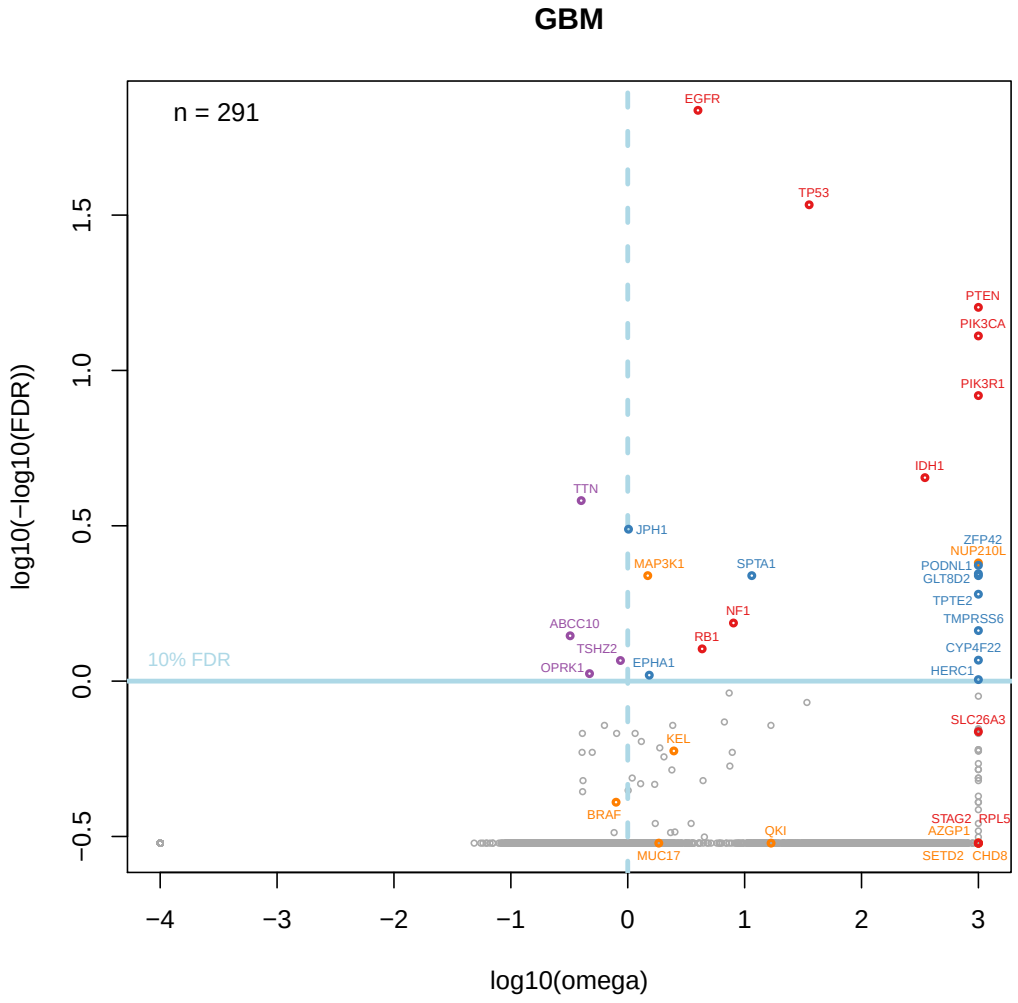


FIGURE 5.8: **Gene-based omega analysis in GBM.** Gene-based PAML results have been displayed in this omega plot for 291 GBM patients to show the omega ratios and FDR values for each gene, together with their level of significance in the Lawrence study. Genes highlighted in red and orange are those shown to be highly significantly mutated (FDR≤0.001) and significantly mutated (FDR≤0.1) respectively in the Lawrence study. Genes highlighted in blue are potential candidate cancer genes shown to reach significance (FDR≤0.1) in the PAML analysis, however do not reach significance in the Lawrence study. Genes highlighted in purple are examples of genes that do not reach significance in the Lawrence study, and have conflicting results in the PAML analysis (with a significant p-value supporting positive selection, but an omega value indicative of negative selection). *R code used to generate plot in Supplementary Appendix D.*

TABLE 5.16: **Ranked list of significant PAML genes in GBM.** The genes found to be significantly mutated in GBM patients in the PAML analysis have been sorted by p-value in ascending order (descending order of significance) in this table. Genes highlighted in red and orange are found to be highly significantly mutated and significantly mutated respectively in the Lawrence study. Genes shown below the blue horizontal line are not found to be significant in the PAML analysis but have been tabulated due to their significance in the Lawrence study. *R and Perl code used to produce list in Supplementary Appendix E.*

Gene	PAML results		Lawrence et al. [2014] p-values			
	Omega	P-value	CV	CL	FN	Combined
EGFR	3.99	4.26e-73	2.43E-15	9.99E-08	9.99E-08	1.11E-16
TP53	35.63	3.41e-38	1.00E-16	9.99E-08	9.99E-08	1.11E-16
PTEN	999.00	7.57e-20	1.00E-16	9.10E-02	4.91E-01	3.55E-15
PIK3CA	999.00	1.13e-16	1.58E-16	9.99E-08	4.66E-02	1.11E-16
PIK3R1	999.00	5.82e-12	5.53E-16	5.53E-06	1.29E-01	1.11E-16
IDH1	348.57	4.22e-08	4.51E-13	4.99E-07	9.16E-01	1.11E-16
JPH1	1.01	1.53e-06	NA	NA	NA	NA
NUP210L	999.00	8.28e-06	5.52E-04	1.05E-01	8.40E-02	9.59E-05
PODNL1	999.00	1.01e-05	NA	NA	NA	NA
ZFP42	999.00	1.55e-05	NA	NA	NA	NA
GLT8D2	999.00	1.83e-05	NA	NA	NA	NA
MAP3K1	1.48	2.10e-05	3.67E-03	5.14E-05	6.35E-02	2.21E-06
SPTA1	11.51	2.12e-05	NA	NA	NA	NA
TPTE2	999.00	4.34e-05	NA	NA	NA	NA
NF1	8.02	1.08e-04	1.00E-16	6.80E-02	8.73E-01	3.55E-15
TMPRSS6	999.00	1.39e-04	NA	NA	NA	NA
RB1	4.33	2.37e-04	1.26E-15	2.33E-01	5.20E-02	2.22E-15
CYP4F22	999.00	3.15e-04	NA	NA	NA	NA
EPHA1	1.53	4.81e-04	NA	NA	NA	NA
HERC1	999.00	5.43e-04	NA	NA	NA	NA
SLC26A3	999.00	1.76e-03	2.62E-07	2.40E-01	4.60E-02	1.90E-07
KEL	2.48	2.88e-03	6.00E-07	2.60E-02	8.80E-01	6.74E-06
BRAF	0.79	6.56e-03	4.99E-02	1.00E-06	1.38E-03	6.85E-07
MUC17	1.84	1.73e-02	1.59E-06	2.71E-01	5.02E-01	1.63E-05
SETD2	999.00	1.15e-01	1.25E-03	1	2.00E-03	7.99E-05
CHD8	999.00	1.26e-01	6.85E-07	1	3.86E-01	7.61E-06
STAG2	999.00	1.54e-01	1.19E-10	1	7.87E-01	2.34E-09
AZGP1	999.00	1.81e-01	3.79E-06	1	7.65E-01	3.56E-05
RPL5	999.00	1.98e-01	1.18E-09	1	3.92E-01	2.07E-08
QKI	16.83	3.32e-01	7.86E-05	1	3.36E-03	2.77E-05
DDX5	NA	NA	6.32E-05	1.65E-02	8.08E-01	1.12E-05
CD1D	NA	NA	5.73E-04	1.00E-03	9.89E-01	7.59E-06

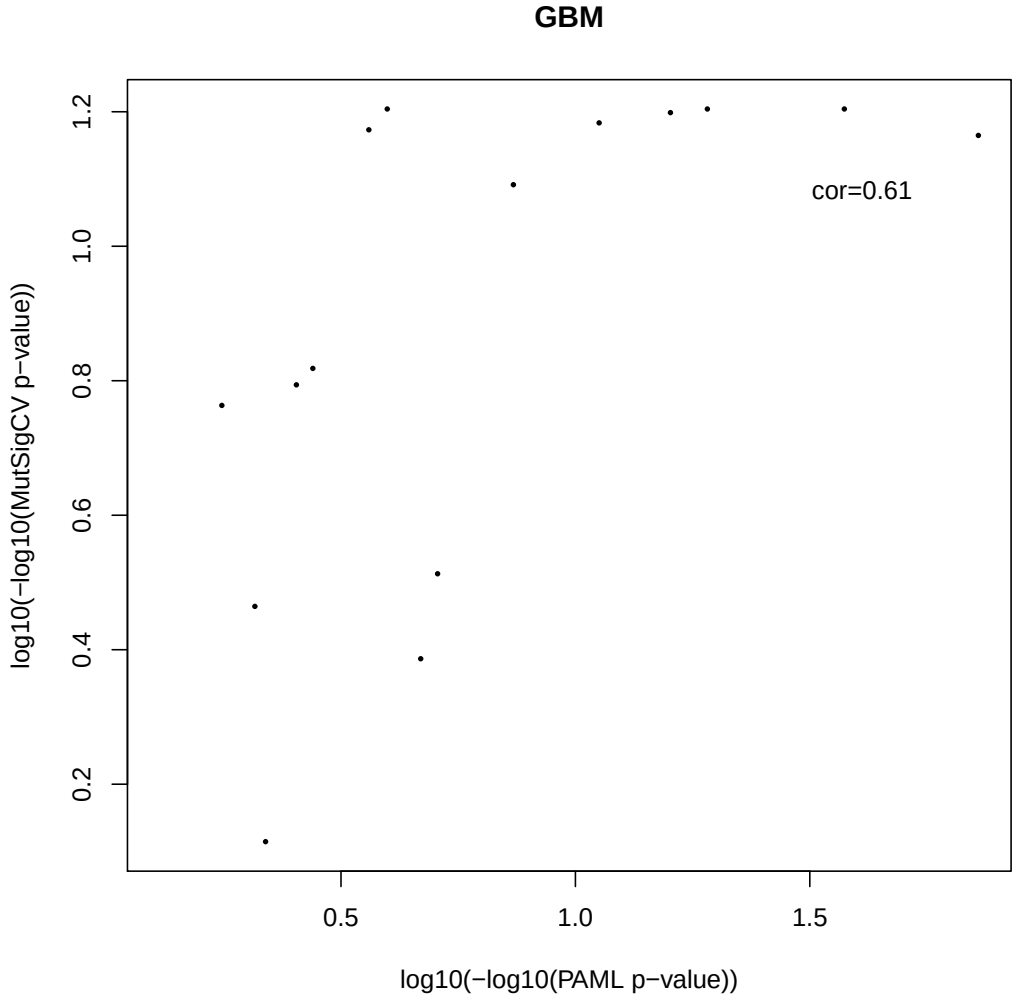


FIGURE 5.9: **Comparison of PAML and MutSig p-values in GBM.** This scatterplot shows the relationship between the p-values obtained from the MutSigCV test in the Lawrence study and the p-values obtained from the PAML analysis in this project, for the GBM subset only. P-values have been double log transformed to show increasing significance with PAML p-value plotted along the x-axis and MutSigCV along the y-axis. Only genes with more significant p-values (<0.05) have been plotted.

the PAML analysis as well as the Lawrence analysis, and they also all contain recurrent mutations (Table 5.17).

TABLE 5.17: **Ranked list of recurrent mutations in GBM.** Ranked list of the top 35 most recurrent SNVs in the GBM subset of the Lawrence dataset, sorted by recurrence in descending order. For each unique mutation, the gene, chromosome, position, ref and alt alleles and how many patients the mutation occurs in (recurrence) have been tabulated.

Gene	Mutation				Recurrence
	Chromosome	Position	Ref	Alt	
IDH1	2	209113112	C	T	13
EGFR	7	55221822	C	T	13
EGFR	7	55233043	G	T	12
TP53	17	7577538	C	T	6
TP53	17	7578406	C	T	5
PTEN	10	89717672	C	T	5
EGFR	7	55221821	G	A	5
BRAF	7	140453136	A	T	5
TP53	17	7577120	C	T	4
TP53	17	7577094	G	A	4
PIK3CA	3	178936091	G	A	4
EGFR	7	55220274	C	T	4
EGFR	7	55211080	G	A	4
C17orf70	17	79516489	C	T	4
ZDHHC4	7	6628405	G	A	3
TSHZ2	20	51870661	G	A	3
TPTE2	13	20039688	G	A	3
TP53	17	7578457	C	T	3
TP53	17	7578263	G	A	3
TP53	17	7578190	T	C	3
TP53	17	7577548	C	T	3
TP53	17	7577539	G	A	3
SPINT1	15	41146113	C	T	3
RB1	13	48953730	C	T	3
PTEN	10	89720852	C	T	3
PTEN	10	89692904	C	T	3
PIK3R1	5	67589138	G	A	3
PCBD2	5	134263015	A	G	3
NUMB	14	73766352	G	T	3
MAP3K1	5	56160697	C	T	3
KLK5	19	51451834	C	T	3
HCN2	19	603979	C	T	3
FTMT	5	121187598	C	T	3
EHD4	15	42193062	G	A	3
EGFR	7	55223543	C	T	3

5.2.6.1 Re-analysis after exclusion of 16 GBM outlier patients

The GBM analysis in PAML was repeated after removing the 16 GBM outlier patients from the GBM dataset. The 16 outlier patients are those that exhibited much higher SNV rates in the TCGA dataset than in the Lawrence dataset, possibly as a result of an increased INDEL rate compared to other patients causing an increase in SNV calls around misaligned reads.

After the exclusion of the 16 outlier patients, the remaining 275 GBM patients were run through PAML which found 22 genes to be significantly mutated (Figure 5.10), compared to 20 in the full GBM dataset of 291 patients (Figure 5.8). The three genes not significantly mutated in the analysis of the whole GBM subset but that reached significance in the re-analysis were DNAH2, MUC17 and TSHz2. CYP4F22 was no longer significantly mutated in the re-analysis, despite being significant in the original GBM PAML analysis.

Of the three genes found to reach significance in the re-analysis, MUC17 was already found to be significantly mutated in the MutSig analysis carried out by [Lawrence et al., 2014] (highlighted in orange in Table 5.18 and Table 5.16). DNAH2 and TSHZ2 however, had not been detected as significantly mutated genes in the Lawrence study. This suggests that by removing the patients with potentially false-positive SNV calls, the PAML analysis is better equipped to detect more of the genes detected by [Lawrence et al., 2014], and has detected two potential candidate genes not previously known to be cancer genes (DNAH2 and TSHZ2).

Figure 5.11 shows the relationship between the PAML and MutSig p-values for the most significant genes, using the PAML p-values calculated in the re-analysis of the 275 GBM patients (excluding the 16 outlier patients). A positive correlation is observed between the two p-values (Pearson's cor = 0.53, p-value = 2.50e-08), however this is a weaker correlation than that previously observed in Figure 5.9 before the removal of the 16 outlier patients. This is an unexpected result, as removing those 16 outlier patients was predicted to remove some of the difference in SNV rates observed between the

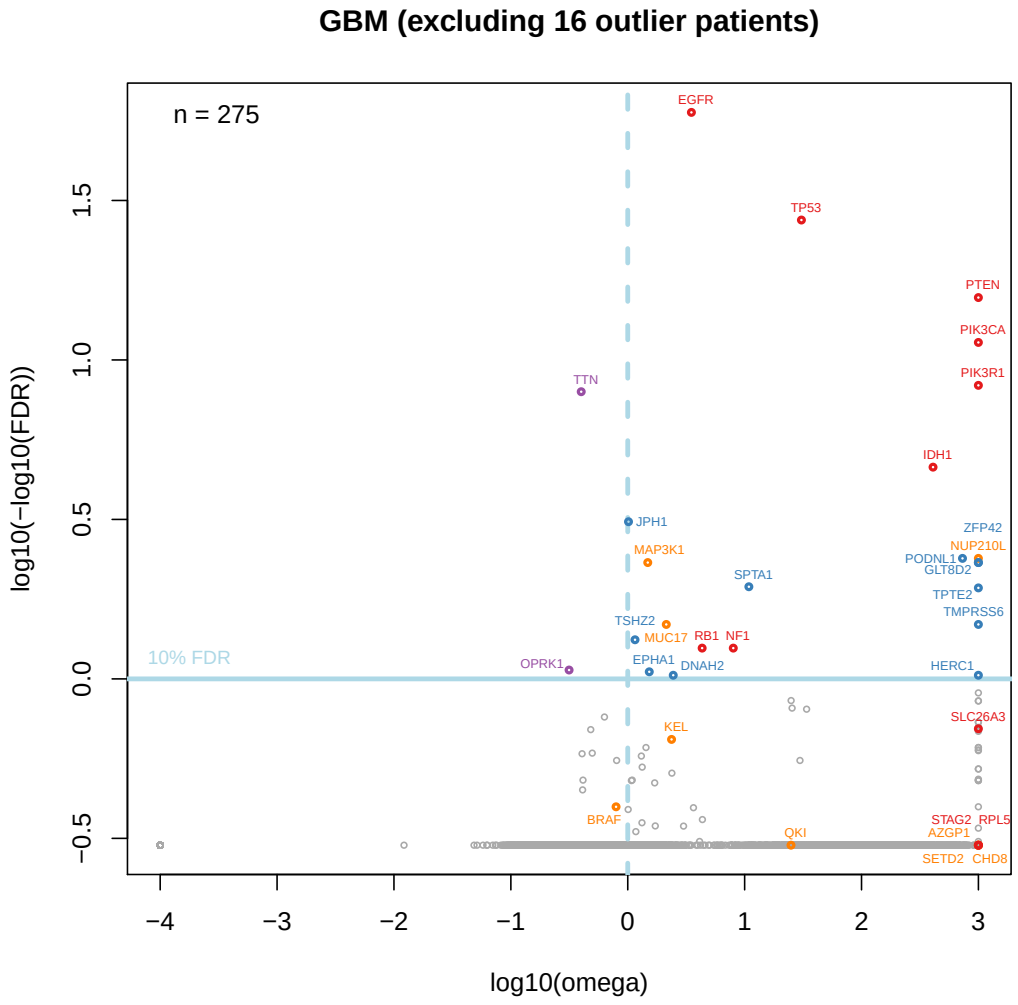


FIGURE 5.10: **Gene-based omega analysis in GBM (excluding 16 outlier patients).** Gene-based PAML results have been displayed in this omega plot for the GBM subset containing 275 patients (exclusive of the 16 outlier patients) to show the omega ratios and FDR values for each gene, together with their level of significance in the Lawrence study. Genes highlighted in red and orange are those shown to be highly significantly mutated (FDR$\leq$0.001) and significantly mutated (FDR$\leq$0.1) respectively in the Lawrence study (over 291 GBM patients). Genes highlighted in blue are potential candidate cancer genes shown to reach significance (FDR$\leq$0.1) in the PAML analysis, however do not reach significance in the Lawrence study. Genes highlighted in purple are examples of genes that do not reach significance in the Lawrence study, and have conflicting results in the PAML analysis (with a significant p-value supporting positive selection, but an omega value indicative of negative selection). *R code used to generate plot in Supplementary Appendix D.*

TABLE 5.18: **Ranked list of significant PAML genes in GBM (excluding 16 outlier patients).** The genes found to be significantly mutated in the subset of 275 GBM patients (exclusive of the 16 outlier patients) in the PAML analysis have been sorted by p-value in ascending order (descending order of significance) in this table. Genes highlighted in red and orange are found to be highly significantly mutated and significantly mutated respectively in the Lawrence study. Genes shown below the blue horizontal line are not found to be significant in the PAML analysis but have been tabulated due to their significance in the Lawrence study of all 291 GBM patients. *R and Perl code used to produce list in Supplementary Appendix E.*

Gene	PAML results		Lawrence et al. [2014] p-values			
	Omega	P-value	CV	CL	FN	Combined
EGFR	3.50	4.69e-64	2.43E-15	9.99E-08	9.99E-08	1.11E-16
TP53	30.60	1.71e-31	1.00E-16	9.99E-08	9.99E-08	1.11E-16
PTEN	999.00	1.52e-19	1.00E-16	9.10E-02	4.91E-01	3.55E-15
PIK3CA	999.00	4.43e-15	1.58E-16	9.99E-08	4.66E-02	1.11E-16
PIK3R1	999.00	5.85e-12	5.53E-16	5.53E-06	1.29E-01	1.11E-16
IDH1	408.97	4.22e-08	4.51E-13	4.99E-07	9.16E-01	1.11E-16
JPH1	1.01	1.53e-06	NA	NA	NA	NA
NUP210L	999.00	9.45e-06	5.52E-04	1.05E-01	8.40E-02	9.59E-05
PODNL1	730.78	1.01e-05	NA	NA	NA	NA
GLT8D2	999.00	1.43e-05	NA	NA	NA	NA
MAP3K1	1.48	1.52e-05	3.67E-03	5.14E-05	6.35E-02	2.21E-06
ZFP42	999.00	1.55e-05	NA	NA	NA	NA
SPTA1	10.87	3.92e-05	NA	NA	NA	NA
TPTE2	999.00	4.36e-05	NA	NA	NA	NA
MUC17	2.14	1.32e-04	1.59E-06	2.71E-01	5.02E-01	1.63E-05
TMPRSS6	999.00	1.39e-04	NA	NA	NA	NA
TSHZ2	1.15	2.10e-04	NA	NA	NA	NA
NF1	8.00	2.67e-04	1.00E-16	6.80E-02	8.73E-01	3.55E-15
RB1	4.33	2.78e-04	1.26E-15	2.33E-01	5.20E-02	2.22E-15
EPHA1	1.53	4.81e-04	NA	NA	NA	NA
HERC1	999.00	5.51e-04	NA	NA	NA	NA
DNAH2	2.45	5.57e-04	NA	NA	NA	NA
SLC26A3	999.00	1.88e-03	2.62E-07	2.40E-01	4.60E-02	1.90E-07
KEL	2.37	2.34e-03	6.00E-07	2.60E-02	8.80E-01	6.74E-06
BRAF	0.79	6.56e-03	4.99E-02	1.00E-06	1.38E-03	6.85E-07
SETD2	999.00	1.13e-01	1.25E-03	1	2.00E-03	7.99E-05
CHD8	999.00	1.26e-01	6.85E-07	1	3.86E-01	7.61E-06
STAG2	999.00	1.73e-01	1.19E-10	1	7.87E-01	2.34E-09
AZGP1	999.00	1.81e-01	3.79E-06	1	7.65E-01	3.56E-05
RPL5	999.00	1.98e-01	1.18E-09	1	3.92E-01	2.07E-08
QKI	24.93	3.32e-01	7.86E-05	1	3.36E-03	2.77E-05
DDX5	NA	NA	6.32E-05	1.65E-02	8.08E-01	1.12E-05
CD1D	NA	NA	5.73E-04	1.00E-03	9.89E-01	7.59E-06

two methods, and hence increase the correlation between the p-value results for both approaches.

Removal of the 16 GBM outlier patients did not significantly affect the recurrence analysis in Table 5.19, suggesting that the SNVs removed were not recurrent mutations, supporting the hypothesis that the high rate of mutations in the 16 outlier patients were mis-called SNVs near INDELs rather than recurrent driver mutations. As seen in the previous GBM analysis including the 16 outlier patients (Table 5.17), 95% (21/22) of the significant mutated genes in the PAML re-analysis were also found to harbour recurrent mutations (same mutation present in more than patient).

Table 5.20 shows the known cancer genes that have been successfully detected in glioblastoma multiforme by both the PAML and recurrence analyses in this project, and also by the Lawrence et al. [2014] MutSig analysis. The same overlap genes were detected in both the GBM analysis including the 16 outliers and in the GBM re-analysis excluding the 16 outliers.

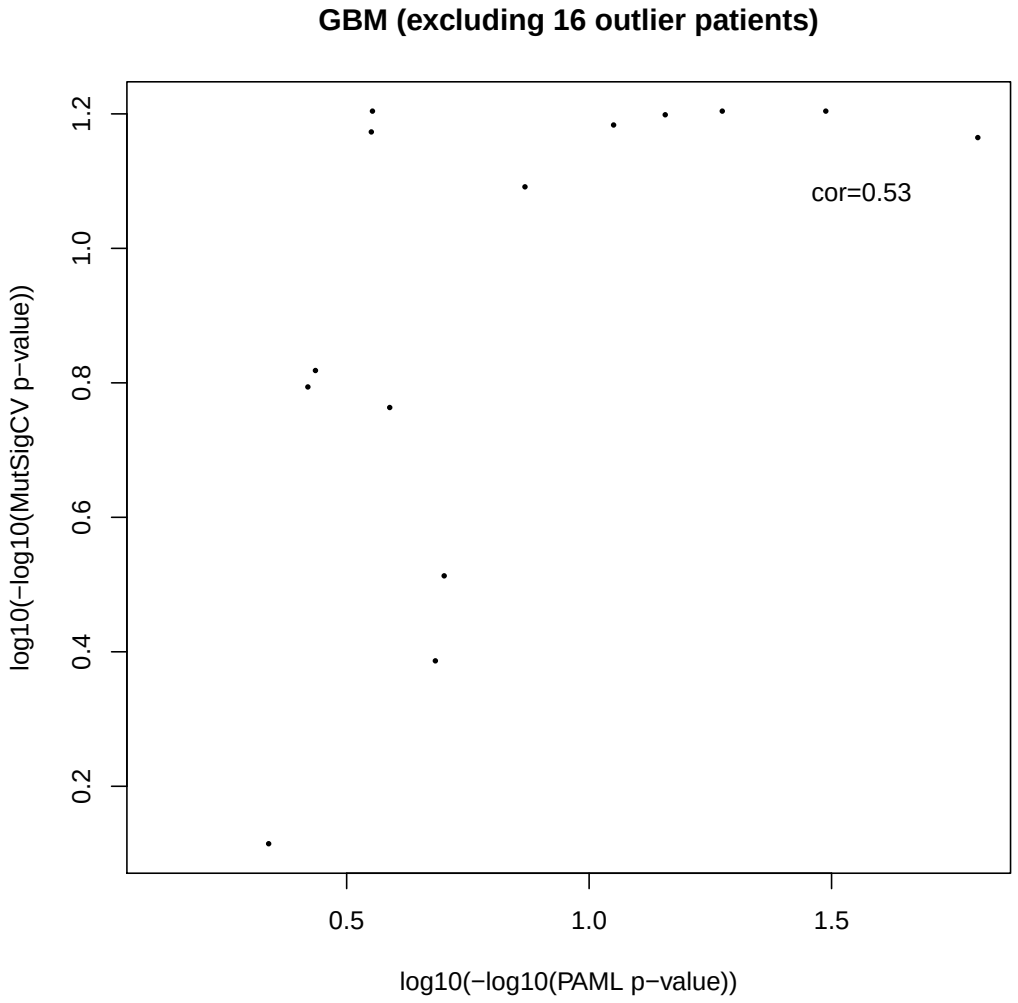


FIGURE 5.11: **Comparison of PAML and MutSig p-values in GBM (excluding 16 outlier patients).** This scatterplot shows the relationship between the p-values obtained from the MutSigCV test in the Lawrence study over all 291 GBM patients, and the p-values obtained from the PAML re-analysis in this project for the 275 GBM patient subset exclusive of the 16 outlier patients. P-values have been double log transformed to show increasing significance with PAML p-value plotted along the x-axis and MutSigCV along the y-axis. Only genes with more significant p-values (<0.05) have been plotted.

TABLE 5.19: **Ranked list of recurrent mutations in GBM (excluding 16 outlier patients).** Ranked list of the top 35 most recurrent SNVs in the GBM subset of the Lawrence dataset containing 275 patients (excluding the 16 outlier patients), sorted by recurrence in descending order. For each unique mutation, the gene, chromosome, position, ref and alt alleles and how many patients the mutation occurs in (recurrence) have been tabulated.

Gene	Mutation				Recurrence
	Chromosome	Position	Ref	Alt	
IDH1	2	209113112	C	T	13
EGFR	7	55221822	C	T	12
EGFR	7	55233043	G	T	11
TP53	17	7577538	C	T	6
TP53	17	7578406	C	T	5
EGFR	7	55221821	G	A	5
BRAF	7	140453136	A	T	5
TP53	17	7577120	C	T	4
PTEN	10	89717672	C	T	4
EGFR	7	55220274	C	T	4
ZDHHC4	7	6628405	G	A	3
TSHZ2	20	51870661	G	A	3
TPTE2	13	20039688	G	A	3
TP53	17	7578457	C	T	3
TP53	17	7578190	T	C	3
TP53	17	7577539	G	A	3
TP53	17	7577094	G	A	3
SPINT1	15	41146113	C	T	3
RB1	13	48953730	C	T	3
PTEN	10	89720852	C	T	3
PTEN	10	89692904	C	T	3
PIK3R1	5	67589138	G	A	3
PIK3CA	3	178936091	G	A	3
PCBD2	5	134263015	A	G	3
NUMB	14	73766352	G	T	3
MAP3K1	5	56160697	C	T	3
KLK5	19	51451834	C	T	3
HCN2	19	603979	C	T	3
FTMT	5	121187598	C	T	3
EHD4	15	42193062	G	A	3
EGFR	7	55223543	C	T	3
EGFR	7	55211080	G	A	3
DST	6	56566691	G	A	3
CNTNAP4	16	76383006	G	T	3
C17orf70	17	79516489	C	T	3

TABLE 5.20: **Cancer gene detection success in glioblastoma multiforme.** Known cancer genes that have been detected as significantly mutated in all three of the aforementioned analyses, overlapping the MutSig analysis in the Lawrence et al. [2014] study, the PAML analysis and the recurrence analysis. The same nine cancer genes were detected in both the set of 291 GBM patients including the 16 outlier patients and in the subset of 275 patients excluding the 16 outlier patients.

Known cancer gene
PTEN
TP53
EGFR
PIK3R1
PIK3CA
NF1
RB1
IDH1
MAP3K1

5.2.7 Head and neck (HNSC)

In the PAML analysis of head and neck cancer, for which 384 patients were analysed, 12 genes were found to be significantly mutated with q≤0.1 and undergoing positive selection with an omega ratio >1 (Figure 5.12).

Amongst the three genes identified as significant in this cancer type by my approach in PAML but not in MutSig, is F-box and WD-40 domain protein 7 (FBXW7). This gene is a known cancer gene previously shown to be associated with colorectal, endometrial cancer and T-cell acute lymphocytic leukemia (T-ALL) [Futreal et al., 2004].

Overall, 30 genes were found to be significantly mutated in at least one of the analyses (Table 5.21). AJUBA was found to be highly significant in the Lawrence study, however PAML has failed to produce results for this gene. The majority of mutations are INDELs in this gene so a significant result would not be expected in PAML, however this does not explain why no results at all have been reported. This could be an issue with the cluster from which PAML was run.

83% (10/12) of the significant PAML genes in HNSC are hit by recurrent mutations. PIK3CA is the most significant gene in the PAML analysis (also highly significant in Lawrence), and is also shown to contain the most recurrent mutation in this cancer type which affects 22 patients in the HNSC dataset (Table 5.22).

Table 5.23 shows the known cancer genes that have been successfully detected in head and neck cancer by both the PAML and recurrence analyses in this project, and also by the Lawrence et al. [2014] MutSig analysis.

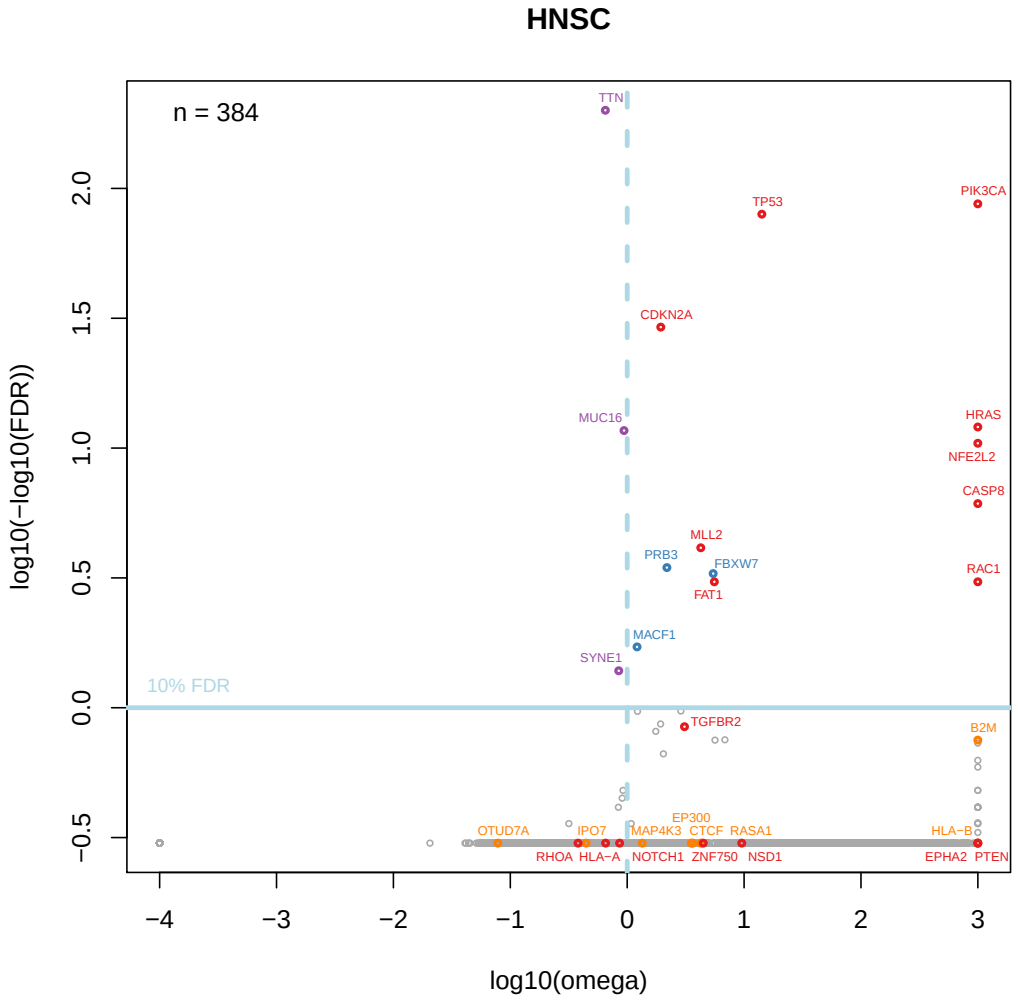


FIGURE 5.12: **Gene-based omega analysis in HNSC.** Gene-based PAML results have been displayed in this omega plot for 384 HNSC patients to show the omega ratios and FDR values for each gene, together with their level of significance in the Lawrence study. Genes highlighted in red and orange are those shown to be highly significantly mutated (FDR≤0.001) and significantly mutated (FDR≤0.1) respectively in the Lawrence study. Genes highlighted in blue are potential candidate cancer genes shown to reach significance (FDR≤0.1) in the PAML analysis, however do not reach significance in the Lawrence study. Genes highlighted in purple are examples of genes that do not reach significance in the Lawrence study, and have conflicting results in the PAML analysis (with a significant p-value supporting positive selection, but an omega value indicative of negative selection). *R code used to generate plot in Supplementary Appendix D.*

TABLE 5.21: **Ranked list of significant PAML genes in HNSC.** The genes found to be significantly mutated in HNSC patients in the PAML analysis have been sorted by p-value in ascending order (descending order of significance) in this table. Genes highlighted in red and orange are found to be highly significantly mutated and significantly mutated respectively in the Lawrence study. Genes shown below the blue horizontal line are not found to be significant in the PAML analysis but have been tabulated due to their significance in the Lawrence study. *R and Perl code used to produce list in Supplementary Appendix E.*

Gene	PAML results		Lawrence et al. [2014] p-values			
	Omega	P-value	CV	CL	FN	Combined
PIK3CA	999.00	1.09e-91	7.39E-07	9.89E-08	3.13E-05	1.98E-12
TP53	14.19	7.02e-84	1.40E-15	5.14E-08	5.14E-08	1.11E-16
CDKN2A	1.94	2.13e-33	8.94E-16	2.76E-03	9.89E-08	1.11E-16
HRAS	999.00	4.01e-16	1.91E-08	7.91E-07	4.42E-04	5.83E-14
NFE2L2	999.00	2.32e-14	2.11E-10	2.27E-06	7.91E-07	7.77E-16
CASP8	999.00	5.61e-10	1.62E-16	3.65E-02	1.57E-02	1.11E-16
MLL2 (KMT2D)	4.26	6.06e-08	1.00E-16	1	1.22E-01	3.55E-15
PRB3	2.19	3.12e-07	NA	NA	NA	NA
FBXW7	5.44	5.18e-07	NA	NA	NA	NA
FAT1	5.57	1.04e-06	4.22E-16	9.78E-01	3.64E-03	3.33E-16
RAC1	999.00	1.04e-06	3.81E-08	3.99E-05	4.61E-01	6.45E-11
MACF1	1.21	2.46e-05	NA	NA	NA	NA
TGFBR2	3.09	2.47e-04	8.40E-08	1.70E-02	5.83E-01	4.08E-08
B2M	999.00	3.58e-04	1.56E-06	1.45E-01	2.52E-01	1.61E-05
EPHA2	999.00	3.09e-03	4.06E-15	5.82E-01	3.53E-01	1.22E-13
NSD1	9.55	4.45e-03	1.00E-16	5.64E-01	3.10E-02	3.55E-15
HLA-A	0.65	6.71e-03	1.45E-06	7.60E-01	6.45E-03	2.46E-07
EP300	3.53	1.50e-02	5.29E-05	1.03E-03	1.12E-01	1.06E-06
ZNF750	4.47	1.64e-02	1.12E-07	2.25E-02	1.91E-03	6.14E-09
PTEN	999.00	4.51e-02	5.06E-09	2.57E-01	4.30E-01	8.10E-08
CTCF	4.26	1.27e-01	3.57E-06	1	1.60E-01	3.38E-05
HLA-B	999.00	1.59e-01	2.71E-06	1	9.60E-02	2.64E-05
RASA1	3.64	2.93e-01	1.36E-05	5.30E-02	1.90E-01	1.10E-04
MAP4K3	1.35	4.81e-01	9.78E-03	2.63E-02	3.92E-02	1.33E-04
RHOA	0.38	5.00e-01	1.62E-04	1.18E-05	3.10E-02	2.18E-08
OTUD7A	0.08	5.00e-01	9.07E-04	2.73E-03	1	2.43E-05
NOTCH1	0.86	5.00e-01	1.00E-16	2.44E-01	4.73E-01	3.55E-15
IPO7	0.45	5.00e-01	1.77E-01	1	1.00E-05	8.42E-05
AJUBA	NA	NA	9.01E-14	1	8.09E-01	2.43E-12

TABLE 5.22: **Ranked list of recurrent mutations in HNSC.** Ranked list of the top 35 most recurrent SNVs in the HNSC subset of the Lawrence dataset, sorted by recurrence in descending order. For each unique mutation, the gene, chromosome, position, ref and alt alleles and how many patients the mutation occurs in (recurrence) have been tabulated.

Gene	Mutation				Recurrence
	Chromosome	Position	Ref	Alt	
PIK3CA	3	178936091	G	A	22
PIK3CA	3	178936082	G	A	15
CDKN2A	9	21971120	G	A	14
TP53	17	7578406	C	T	9
CDKN2A	9	21971186	G	A	9
TP53	17	7578212	G	A	8
TP53	17	7577538	C	T	8
TP53	17	7577094	G	A	8
PIK3CA	3	178952085	A	G	8
TP53	17	7577120	C	T	7
TP53	17	7578263	G	A	6
TP53	17	7577548	C	T	6
TP53	17	7577539	G	A	6
CDKN2A	9	21971028	C	T	6
CDKN2A	9	21971000	C	A	6
RHOA	3	49412905	C	G	5
CDKN2A	9	21971029	C	T	5
TP53	17	7578395	G	A	4
TP53	17	7577022	G	A	4
NFE2L2	2	178098810	C	G	4
MB21D2	3	192516720	G	C	4
HRAS	11	534285	C	A	4
EP300	22	41565529	G	A	4
TP53	17	7579358	C	A	3
TP53	17	7578524	G	A	3
TP53	17	7578461	C	A	3
TP53	17	7578394	T	C	3
TP53	17	7578271	T	A	3
TP53	17	7578190	T	C	3
TP53	17	7577556	C	A	3
TP53	17	7577106	G	A	3
TP53	17	7577046	C	A	3
PRB3	12	11421068	G	A	3
NFE2L2	2	178098960	C	G	3
MAPK1	22	22127164	C	T	3

TABLE 5.23: **Cancer gene detection success in head and neck cancer.** Known cancer genes that have been detected as significantly mutated in all three of the aforementioned analyses, overlapping the MutSig analysis in the Lawrence et al. [2014] study, the PAML analysis and the recurrence analysis.

Known cancer gene
TP53
CDKN2A
FAT1
PIK3CA
CASP8
NFE2L2
HRAS
RAC1

5.2.8 Kidney clear cell (KIRC)

KIRC is the most common form of kidney cancer [Kaelin, 2010].

In the PAML analysis of kidney clear cell cancer, for which 417 patients were analysed, 12 genes were found to be significantly mutated with q≤0.1 and undergoing positive selection with an omega ratio >1 (Figure 5.13).

VHL is the most significant result in the PAML analysis as would be expected, since this gene is mutated or silenced in over 50% of sporadic kidney renal clear cell carcinomas and is known to be causally implicated in both sporadic and germline kidney renal clear cell carcinomas [Kaelin, 2004].

Phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha (PIK3CA) is a known cancer gene that has been detected by the PAML analysis as a significantly mutated gene in this cancer type, however it has not been detected by the MutSig analysis. PIK3CA is associated with colorectal, gastric, glioblastoma and breast cancers [Futreal et al., 2004].

Overall 20 genes were found to be significant in at least one of the analyses (Table 5.24).

75% (9/12) of the significant PAML genes in KIRC are hit by recurrent mutations. As expected VHL contains many recurrent mutations across all KIRC patients, and is the most recurrently mutated gene in KIRC (Table 5.25).

Table 5.26 shows the known cancer genes that have been successfully detected in kidney clear cell cancer by both the PAML and recurrence analyses in this project, and also by the Lawrence et al. [2014] MutSig analysis.

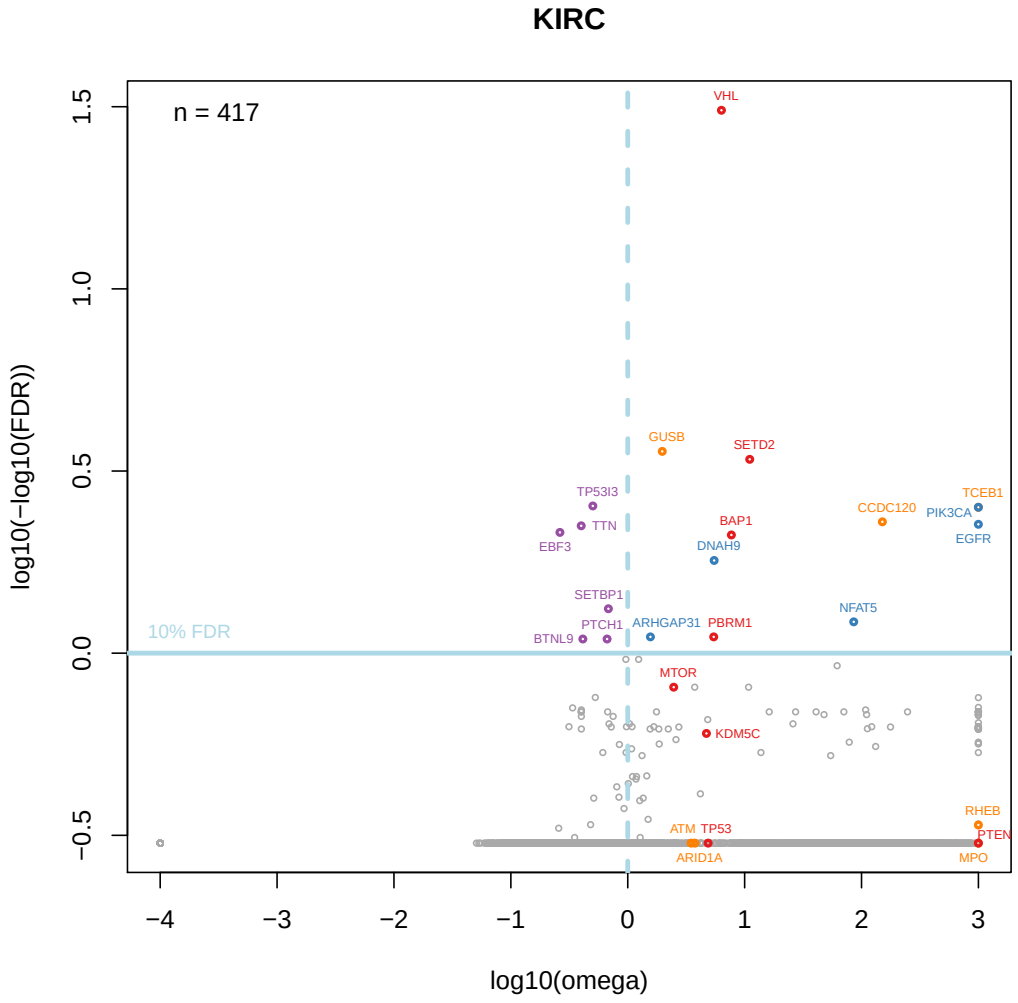


FIGURE 5.13: **Gene-based omega analysis in KIRC.** Gene-based PAML results have been displayed in this omega plot for 417 KIRC patients to show the omega ratios and FDR values for each gene, together with their level of significance in the Lawrence study. Genes highlighted in red and orange are those shown to be highly significantly mutated (FDR≤0.001) and significantly mutated (FDR≤0.1) respectively in the Lawrence study. Genes highlighted in blue are potential candidate cancer genes shown to reach significance (FDR≤0.1) in the PAML analysis, however do not reach significance in the Lawrence study. Genes highlighted in purple are examples of genes that do not reach significance in the Lawrence study, and have conflicting results in the PAML analysis (with a significant p-value supporting positive selection, but an omega value indicative of negative selection). *R code used to generate plot in Supplementary Appendix D.*

TABLE 5.24: **Ranked list of significant PAML genes in KIRC.** The genes found to be significantly mutated in KIRC patients in the PAML analysis have been sorted by p-value in ascending order (descending order of significance) in this table. Genes highlighted in red and orange are found to be highly significantly mutated and significantly mutated respectively in the Lawrence study. Genes shown below the blue horizontal line are not found to be significant in the PAML analysis but have been tabulated due to their significance in the Lawrence study. *R and Perl code used to produce list in Supplementary Appendix E.*

Gene	PAML results		Lawrence et al. [2014] p-values			
	Omega	P-value	CV	CL	FN	Combined
VHL	6.33	2.25e-35	1.00E-16	9.84E-01	9.31E-01	3.55E-15
GUSB	1.97	9.91e-08	1.13E-01	1.00E-06	8.15E-01	1.07E-05
SETD2	11.05	2.23e-07	1.63E-15	6.81E-03	5.38E-03	2.22E-16
PIK3CA	999.00	2.89e-06	NA	NA	NA	NA
TCEB1	999.00	3.46e-06	4.45E-04	4.82E-03	5.37E-01	2.39E-05
CCDC120	150.76	6.70e-06	4.76E-02	1.00E-04	4.58E-01	4.36E-05
EGFR	999.00	8.32e-06	NA	NA	NA	NA
BAP1	7.70	1.61e-05	1.00E-16	6.64E-03	2.59E-01	2.22E-16
DNAH9	5.48	3.61e-05	NA	NA	NA	NA
NFAT5	85.79	1.59e-04	NA	NA	NA	NA
ARHGAP31	1.56	2.35e-04	NA	NA	NA	NA
PBRM1	5.44	2.35e-04	1.00E-16	1	7.90E-02	3.55E-15
MTOR	2.47	7.06e-04	5.27E-05	1.52E-05	8.07E-01	2.05E-08
KDM5C	4.71	3.96e-03	1.97E-16	4.67E-01	4.98E-01	6.55E-15
RHEB	999.00	9.97e-03	5.74E-03	7.06E-04	2.29E-01	2.13E-05
PTEN	999.00	6.25e-02	1.00E-16	2.45E-01	6.88E-01	3.55E-15
TP53	4.87	8.35e-02	9.53E-12	1	2.50E-02	2.12E-10
MPO	999.00	2.13e-01	2.90E-02	1	1.00E-04	2.80E-05
ATM	3.48	5.00e-01	9.71E-04	1.05E-01	4.27E-03	2.19E-05
ARID1A	3.75	5.00e-01	6.38E-07	1	1.53E-01	7.13E-06

TABLE 5.25: **Ranked list of recurrent mutations in KIRC.** Ranked list of the top 35 most recurrent SNVs in the KIRC subset of the Lawrence dataset, sorted by recurrence in descending order. For each unique mutation, the gene, chromosome, position, ref and alt alleles and how many patients the mutation occurs in (recurrence) have been tabulated.

Gene	Mutation				Recurrence
	Chromosome	Position	Ref	Alt	
VHL	3	10183797	T	A	7
VHL	3	10183725	C	A	7
VHL	3	10183863	G	A	5
VHL	3	10183734	C	A	5
VHL	3	10191572	G	T	3
VHL	3	10191558	T	C	3
VHL	3	10191513	T	C	3
VHL	3	10191479	C	G	3
VHL	3	10191469	A	G	3
VHL	3	10188321	G	T	3
VHL	3	10188240	T	A	3
VHL	3	10183817	C	T	3
VHL	3	10183794	G	A	3
VHL	3	10183776	G	C	3
VHL	3	10183752	T	A	3
VHL	3	10183748	C	T	3
VHL	3	10183665	C	T	3
PIK3CA	3	178936091	G	A	3
PBRM1	3	52621464	G	A	3
LIN28A	1	26742193	A	C	3
ZNF780B	19	40540306	G	A	2
ZNF732	4	265967	C	G	2
ZNF687	1	151261955	T	G	2
ZNF462	9	109688141	T	A	2
ZNF43	19	21992281	A	C	2
ZNF292	6	87970404	G	T	2
ZNF273	7	64388663	T	A	2
ZMAT5	22	30134342	C	A	2
ZFYVE9	1	52703736	T	A	2
ZFR2	19	3833722	A	T	2
ZFAND2B	2	220072989	T	C	2
ZDHHC1	16	67428940	T	A	2
WWP2	16	69964089	G	C	2
VHL	3	10191570	T	C	2
VHL	3	10191488	C	T	2

TABLE 5.26: **Cancer gene detection success in kidney clear cell cancer.** Known cancer genes that have been detected as significantly mutated in all three of the aforementioned analyses, overlapping the MutSig analysis in the Lawrence et al. [2014] study, the PAML analysis and the recurrence analysis.

Known cancer gene
VHL
PBRM1
BAP1

5.2.9 Lung adenocarcinoma (LUAD)

In the PAML analysis of lung adenocarcinoma, for which 400 patients were analysed, 15 genes were found to be significantly mutated with q≤0.1 and undergoing positive selection with an omega ratio >1 (Figure 5.14).

Catenin (cadherin-associated protein) beta 1 (CTNNB1) has been detected as significantly mutated in this cancer type by PAML but not by MutSig, and is a known cancer gene associated with colorectal, ovarian, hepatoblastoma and pleomorphic salivary gland adenoma amongst other tumour types [Futreal et al., 2004].

Overall, 30 genes were found to be significant in at least one of the analyses (Table 5.27). The top three hits, XIRP2, FLG and MUC16 had FDR values of infinity as their p-values were 0, therefore these genes have come up as highly significant in the PAML analysis. However, Lawrence has not detected these as significant. It is possible that these are false-positive results, especially MUC16 since this gene is known to appear as a false-positive in similar such cancer studies [Lawrence et al., 2013]. FLG generally has a very high mutation rate, caused by the presence of highly mutable long simple tandem repeats (STRs) in this gene, and has been linked to autoinflammatory diseases [Ross, 2014].

79% (11/14) of the significantly mutated PAML genes in LUAD are hit by recurrent mutation (Table 5.28). XIRP2 is not hit by recurrent mutations in this cancer. This supports the speculation that it may be a false-positive.

Table 5.29 shows the known cancer genes that have been successfully detected in lung adenocarcinoma by both the PAML and recurrence analyses in this project, and also by the Lawrence et al. [2014] MutSig analysis.

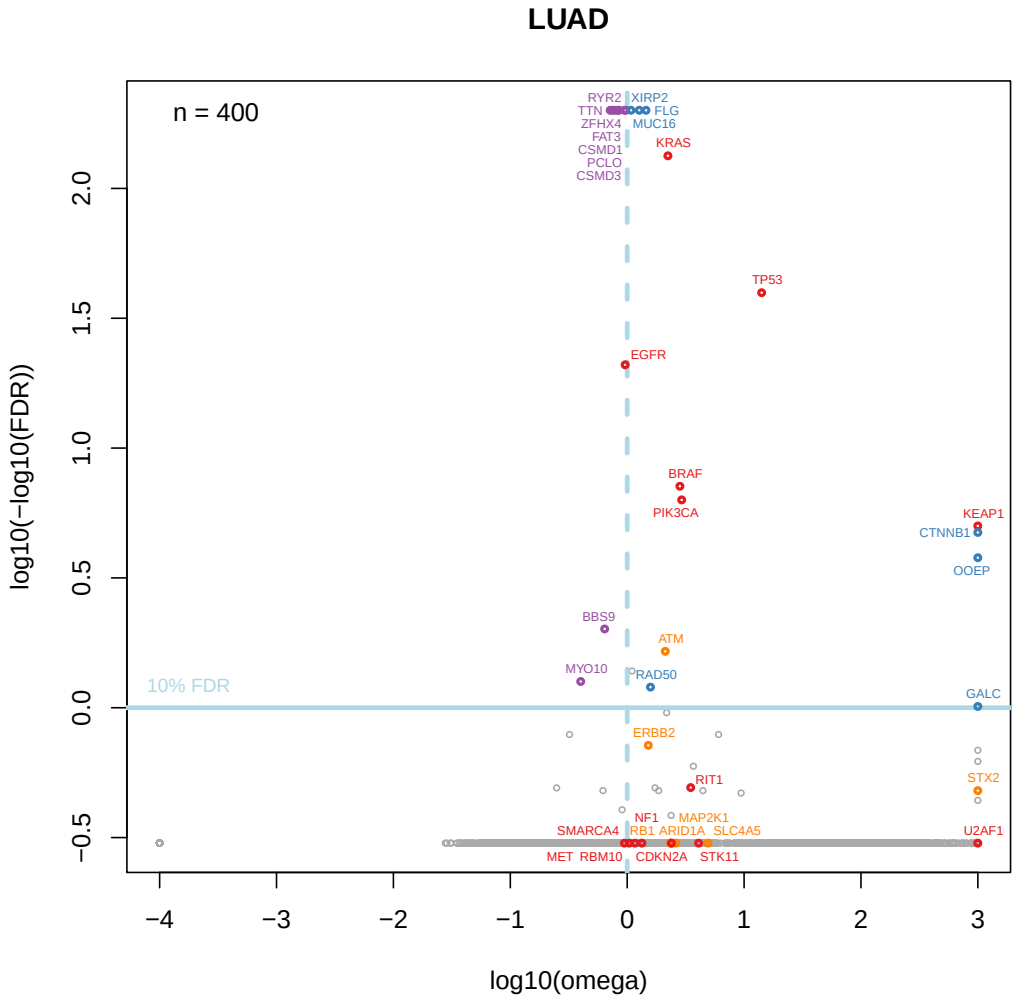


FIGURE 5.14: **Gene-based omega analysis in LUAD.** Gene-based PAML results have been displayed in this omega plot for 400 LUAD patients to show the omega ratios and FDR values for each gene, together with their level of significance in the Lawrence study. Genes highlighted in red and orange are those shown to be highly significantly mutated (FDR≤0.001) and significantly mutated (FDR≤0.1) respectively in the Lawrence study. Genes highlighted in blue are potential candidate cancer genes shown to reach significance (FDR≤0.1) in the PAML analysis, however do not reach significance in the Lawrence study. Genes highlighted in purple are examples of genes that do not reach significance in the Lawrence study, and have conflicting results in the PAML analysis (with a significant p-value supporting positive selection, but an omega value indicative of negative selection). *R code used to generate plot in Supplementary Appendix D.*

TABLE 5.27: **Ranked list of significant PAML genes in LUAD.** The genes found to be significantly mutated in LUAD patients in the PAML analysis have been sorted by p-value in ascending order (descending order of significance) in this table. Genes highlighted in red and orange are found to be highly significantly mutated and significantly mutated respectively in the Lawrence study. Genes shown below the blue horizontal line are not found to be significant in the PAML analysis but have been tabulated due to their significance in the Lawrence study. *R and Perl code used to produce list in Supplementary Appendix E.*

Gene	PAML results		Lawrence et al. [2014] p-values			
	Omega	P-value	CV	CL	FN	Combined
FLG	1.45	0.00e+00	NA	NA	NA	NA
MUC16	1.08	0.00e+00	NA	NA	NA	NA
XIRP2	1.27	0.00e+00	NA	NA	NA	NA
KRAS	2.23	2.16e-137	3.91E-08	9.89E-08	9.89E-08	1.16E-13
TP53	14.13	1.77e-43	4.02E-15	1.62E-04	5.04E-08	1.11E-16
EGFR	0.96	1.04e-24	2.58E-06	9.89E-08	2.08E-03	6.61E-12
BRAF	2.82	7.11e-11	3.55E-03	5.93E-07	5.86E-01	6.29E-08
PIK3CA	2.92	4.95e-10	3.71E-02	3.96E-07	4.26E-02	2.70E-07
KEAP1	999.00	1.06e-08	4.12E-15	3.52E-01	1.36E-05	1.11E-16
CTNNB1	999.00	2.15e-08	NA	NA	NA	NA
OOEP	999.00	2.03e-07	NA	NA	NA	NA
ATM	2.12	3.07e-05	4.97E-03	1.44E-02	1.42E-02	2.60E-05
NBPF15	1.10	5.90e-05	NA	NA	NA	NA
RAD50	1.58	9.90e-05	NA	NA	NA	NA
GALC	999.00	1.60e-04	NA	NA	NA	NA
ERBB2	1.52	3.68e-04	7.80E-03	6.50E-06	3.94E-01	1.32E-06
RIT1	3.49	7.03e-04	5.27E-07	1.22E-02	1.36E-01	4.17E-08
STX2	999.00	8.51e-04	3.42E-03	1.27E-02	2.23E-02	3.93E-05
MAP2K1	2.61	3.26e-03	4.72E-04	1.11E-02	8.89E-01	4.95E-05
STK11	4.08	3.61e-03	1.00E-16	9.87E-01	7.00E-02	3.55E-15
SLC4A5	4.91	1.85e-02	2.58E-04	1.55E-02	1.47E-01	3.19E-05
CDKN2A	1.34	3.79e-02	3.70E-09	1.49E-01	9.89E-08	1.19E-14
U2AF1	999.00	9.70e-02	2.87E-07	7.71E-06	2.83E-05	7.98E-13
MET	0.95	1.09e-01	2.44E-06	1.03E-03	2.15E-03	2.64E-09
RB1	2.34	3.37e-01	9.01E-06	1	3.61E-01	7.68E-05
SMARCA4	1.17	4.75e-01	5.88E-14	1	2.63E-01	1.61E-12
RBM10	2.39	5.00e-01	3.58E-08	1	1.91E-01	5.03E-07
ARID1A	2.41	5.00e-01	5.02E-07	3.73E-01	5.43E-01	5.73E-06
NF1	1.04	5.00e-01	2.20E-10	6.50E-01	4.30E-02	4.21E-09
NBPF1	NA	NA	4.94E-01	2.94E-05	9.02E-03	3.25E-05

TABLE 5.28: **Ranked list of recurrent mutations in LUAD.** Ranked list of the top 35 most recurrent SNVs in the LUAD subset of the Lawrence dataset, sorted by recurrence in descending order. For each unique mutation, the gene, chromosome, position, ref and alt alleles and how many patients the mutation occurs in (recurrence) have been tabulated.

Gene	Mutation				Recurrence
	Chromosome	Position	Ref	Alt	
KRAS	12	25398285	C	A	55
KRAS	12	25398284	C	A	20
EGFR	7	55259515	T	G	13
U2AF1	21	44524456	G	A	10
KRAS	12	25398284	C	G	8
TP53	17	7578457	C	A	7
TP53	17	7577120	C	A	6
PIK3CA	3	178936091	G	A	6
KRAS	12	25398284	C	T	6
KRAS	12	25398281	C	T	5
KRAS	12	25398282	C	A	4
EGFR	7	55241708	G	C	4
CTNNB1	3	41266113	C	T	4
BRAF	7	140481402	C	A	4
TTN	2	179553413	G	T	3
TP53	17	7579377	G	A	3
TP53	17	7579312	C	A	3
TP53	17	7578500	G	A	3
TP53	17	7578469	C	A	3
TP53	17	7578461	C	A	3
TP53	17	7578455	C	G	3
TP53	17	7578442	T	C	3
TP53	17	7578406	C	T	3
TP53	17	7578263	G	A	3
TP53	17	7578212	G	A	3
TP53	17	7577538	C	G	3
TP53	17	7577536	T	C	3
TP53	17	7577535	C	A	3
TP53	17	7577114	C	A	3
TP53	17	7574000	C	A	3
TBC1D12	10	96162370	G	A	3
SLC6A17	1	110734898	G	T	3
RAD50	5	131895051	G	T	3
PI15	8	75737548	G	T	3
NRAS	1	115256529	T	A	3

TABLE 5.29: **Cancer gene detection success in lung adenocarcinoma.** Known cancer genes that have been detected as significantly mutated in all three of the aforementioned analyses, overlapping the MutSig analysis in the Lawrence et al. [2014] study, the PAML analysis and the recurrence analysis.

Known cancer gene
TP53
KRAS
KEAP1
BRAF
PIK3CA

5.2.10 Lung squamous cell carcinoma (LUSC)

In the PAML analysis of lung squamous cell carcinoma, for which 177 patients were analysed, eight genes were found to be significantly mutated with q≤0.1 and undergoing positive selection with an omega ratio >1 (Figure 5.15).

DMD is the only gene that PAML has found to be significant but that did not reach significance in Lawrence. DMD encodes the protein dystrophin found in cardiac and skeletal muscle. It is a highly complex gene and is the largest known human gene encompassing 2.6 million base pairs of DNA and containing 79 exons [Nowak and Davies, 2004]. Both DMD-associated dilated cardiomyopathy and Duchenne and Becker muscular dystrophy are caused by mutations in this gene, however it has not been identified as a cancer causing gene in the Cancer Gene Census [Futreal et al., 2004]. Recently however, Wang et al. [2014] validated DMD as a tumour suppressor in common human mesenchymal tumors featuring myogenic differentiation, including gastrointestinal stromal tumor (GIST), rhabdomyosarcoma (RMS) and leiomyosarcoma (LMS), and likely anti-metastatic factor. An intragenic deletion was found to be a frequent mechanism by which myogenic tumours progress to high-grade, lethal sarcomas. These results suggest that therapies in development for muscular dystrophies may also have relevance in the treatment of cancer. Although lung squamous cell carcinoma is a cancer of the epithelial cells rather than a myogenic cancer affecting the muscular tissue, the fact that this gene has now been validated as a tumour suppressor gene is promising for its candidature as a cancer gene in other cancer types.

Overall, 14 genes were found to reach significance in at least one of the two analyses (Table 5.30).

Of the significantly mutated PAML genes, 67% (6/9) contain recurrent mutations including PIK3CA and TP53 (Table 5.31).

Table 5.32 shows the known cancer genes that have been successfully detected in lung squamous cell carcinoma by both the PAML and recurrence analyses in this project, and also by the Lawrence et al. [2014] MutSig analysis.

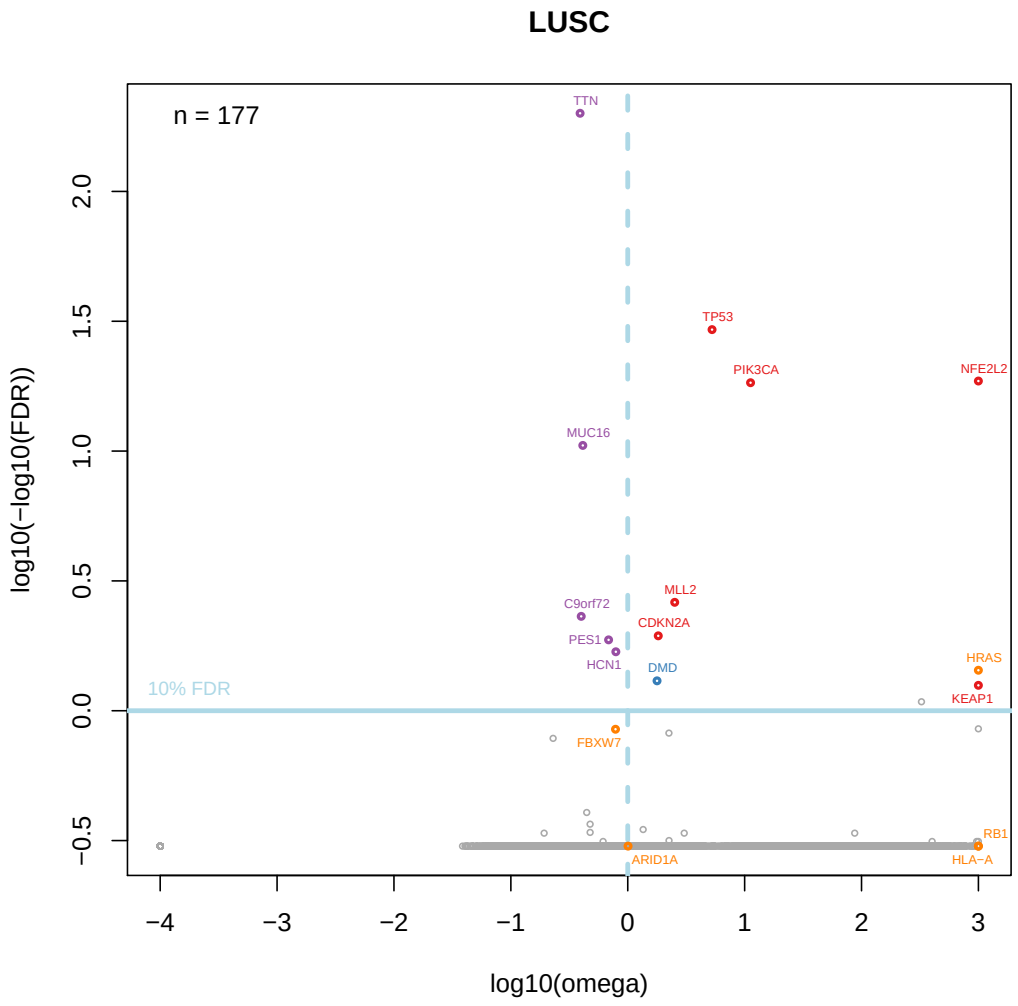


FIGURE 5.15: **Gene-based omega analysis in LUSC.** Gene-based PAML results have been displayed in this omega plot for 177 LUSC patients to show the omega ratios and FDR values for each gene, together with their level of significance in the Lawrence study. Genes highlighted in red and orange are those shown to be highly significantly mutated (FDR≤0.001) and significantly mutated (FDR≤0.1) respectively in the Lawrence study. Genes highlighted in blue are potential candidate cancer genes shown to reach significance (FDR≤0.1) in the PAML analysis, however do not reach significance in the Lawrence study. Genes highlighted in purple are examples of genes that do not reach significance in the Lawrence study, and have conflicting results in the PAML analysis (with a significant p-value supporting positive selection, but an omega value indicative of negative selection). *R code used to generate plot in Supplementary Appendix D.*

TABLE 5.30: **Ranked list of significant PAML genes in LUSC.** The genes found to be significantly mutated in LUSC patients in the PAML analysis have been sorted by p-value in ascending order (descending order of significance) in this table. Genes highlighted in red and orange are found to be highly significantly mutated and significantly mutated respectively in the Lawrence study. Genes shown below the blue horizontal line are not found to be significant in the PAML analysis but have been tabulated due to their significance in the Lawrence study. *R and Perl code used to produce list in Supplementary Appendix E.*

Gene	PAML results		Lawrence et al. [2014] p-values			
	Omega	P-value	CV	CL	FN	Combined
TP53	5.26	8.14e-34	1.00E-16	9.89E-08	9.89E-08	1.11E-16
NFE2L2	999.00	6.89e-23	1.93E-02	9.89E-08	9.89E-08	3.24E-08
PIK3CA	11.23	1.73e-22	6.91E-05	9.89E-08	1.86E-02	1.55E-10
MLL2 (KMT2D)	2.52	1.37e-06	1.67E-12	1	4.68E-01	4.02E-11
CDKN2A	1.82	8.59e-06	1.00E-16	2.63E-01	9.89E-08	1.11E-16
HRAS	999.00	3.83e-05	6.90E-04	3.82E-03	1.61E-02	1.38E-05
DMD	1.78	5.61e-05	NA	NA	NA	NA
KEAP1	999.00	6.84e-05	9.69E-11	3.80E-02	1.62E-01	8.56E-11
FBXW7	0.79	2.13e-04	2.89E-03	8.27E-03	5.45E-02	5.61E-05
RB1	999.00	1.36e-01	6.16E-08	1	7.34E-01	8.32E-07
HLA-A	999.00	1.87e-01	7.06E-05	1	7.73E-03	1.83E-05
ARID1A	1.01	5.00e-01	2.68E-06	7.70E-02	8.88E-01	2.61E-05

TABLE 5.31: **Ranked list of recurrent mutations in LUSC.** Ranked list of the top 35 most recurrent SNVs in the LUSC subset of the Lawrence dataset, sorted by recurrence in descending order. For each unique mutation, the gene, chromosome, position, ref and alt alleles and how many patients the mutation occurs in (recurrence) have been tabulated.

Gene	Mutation				Recurrence
	Chromosome	Position	Ref	Alt	
PIK3CA	3	178936091	G	A	10
SUMF1	3	3940376	G	T	6
TP53	17	7578457	C	A	5
RAPGEF4	2	173740429	T	C	5
PTPRD	9	9372392	T	G	5
ZNF566	19	36982883	T	C	4
TPO	2	1461103	A	G	4
TP53	17	7579312	C	A	4
SSFA2	2	182790804	C	T	4
SLCO1B3	12	21239914	T	C	4
SLC9A7	23	46520751	T	C	4
PDSS2	6	107486359	G	C	4
OTOA	16	21766513	A	G	4
NFE2L2	2	178098810	C	G	4
NEK7	1	198270426	G	T	4
GSTA1	6	52669326	C	A	4
FIG4	6	110075002	G	A	4
DNAH14	1	225164811	C	G	4
DNAH14	1	225164801	C	T	4
DIP2C	10	393257	A	C	4
CDH16	16	66954690	C	T	4
CD2AP	6	47557303	T	C	4
C7orf31	7	25177201	C	T	4
ATP8A2	13	26470951	T	C	4
AFF2	23	147847963	A	T	4
ZNF804B	7	88645165	C	G	3
ZNF717	3	75834823	T	A	3
ZNF343	20	2483188	A	G	3
ZNF33B	10	43123392	G	A	3
ZGPAT	20	62354858	T	C	3
ZFP41	8	144342202	G	C	3
WDR20	14	102638729	C	T	3
URB1	21	33762766	C	T	3
UBR1	15	43302963	G	C	3
UBE2H	7	129581039	C	T	3

TABLE 5.32: **Cancer gene detection success in lung squamous cell carcinoma.** Known cancer genes that have been detected as significantly mutated in all three of the aforementioned analyses, overlapping the MutSig analysis in the Lawrence et al. [2014] study, the PAML analysis and the recurrence analysis.

Known cancer gene
TP53
CDKN2A
PIK3CA
NFE2L2
KEAP1

5.2.11 Melanoma (MEL)

In the PAML analysis of melanoma, for which 118 patients were analysed, 105 genes were found to be significantly mutated with q≤0.1 and undergoing positive selection with an omega ratio >1 (Figure 5.16).

NRAS and BRAF were found to be the most significantly mutated genes in the PAML analysis, and were also found to be highly significant in the Lawrence study. This result is expected since these two genes have both been causally implicated in melanoma. BRAF is a serine/threonine kinase that is commonly activated by somatic point mutation in human cancer. About 50% of melanomas harbour activating BRAF mutations, with over 90% of these at V600E [Ascierto et al., 2012]. NRAS is an N-ras oncogene and is mutated in approximately 20% of melanomas [Atefi et al., 2015].

F-box and WD-40 domain protein 7 (FBXW7) was amongst the genes detected as significantly mutated in melanoma only in the PAML analysis. This gene is a known cancer gene previously shown to be associated with colorectal, endometrial cancer and T-cell acute lymphocytic leukemia (T-ALL) [Futreal et al., 2004].

Many genes are mutated in melanoma due to the high mutation rate observed in this cancer type. Overall, 113 genes were found to be significantly mutated in at least one of the analyses (Table 5.33).

88% (91/103) of the significantly mutated PAML genes for MEL are hit by recurrent mutations. NRAS and BRAF, as expected, are amongst those affected by recurrent mutations (Table 5.34). Somatic activating mutations in BRAF have been found in approximately 50% of melanomas [Ascierto et al., 2012]. In this Lawrence dataset of 118 melanoma patients, BRAF is shown to be hit recurrently by the same mutation in 62 patients; this is the V600E mutation (T→A (A→T on complementary strand) single nucleotide substitution at codon 600 (genomic position 140453136) resulting in a valine to glutamic acid change) that has been shown to be the most common BRAF mutation in melanoma, accounting for over 81% of all BRAF mutations in melanoma [Ascierto et al., 2012].

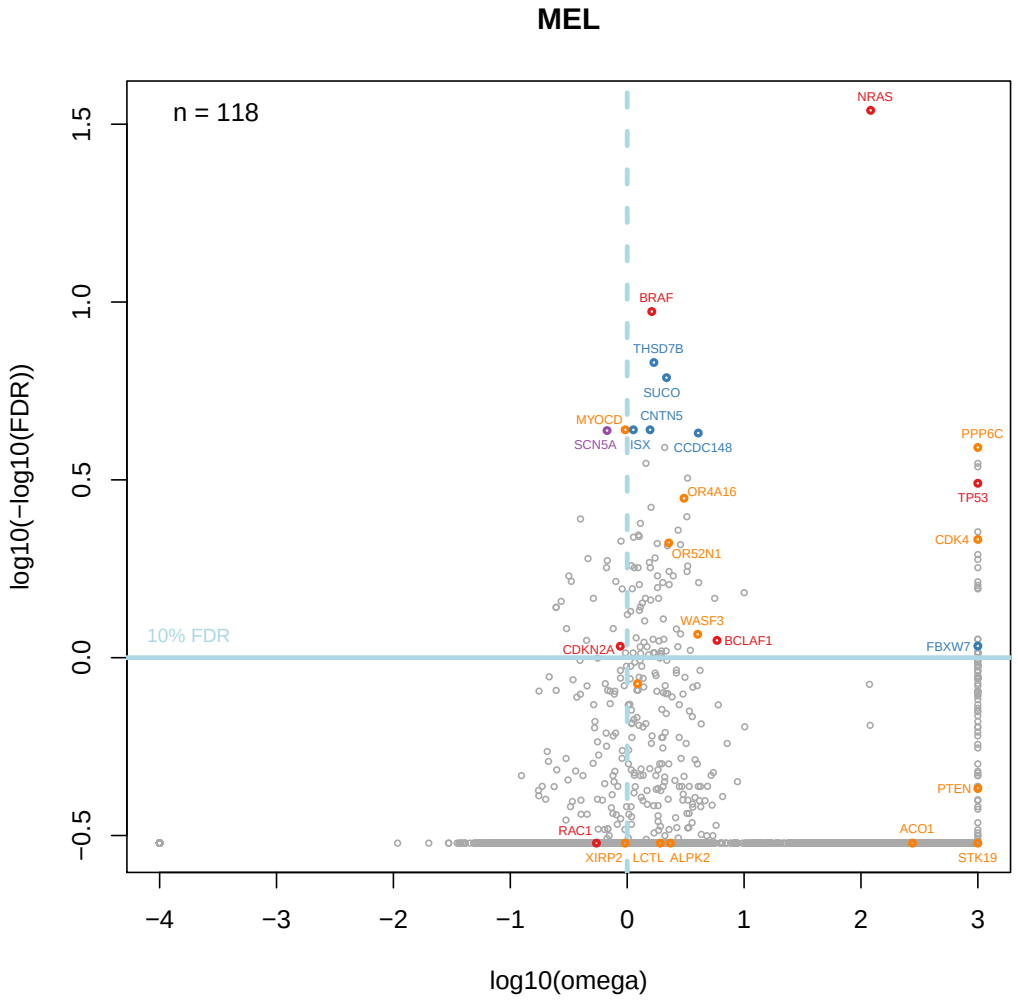


FIGURE 5.16: **Gene-based omega analysis in MEL.** Gene-based PAML results have been displayed in this omega plot for 118 MEL patients to show the omega ratios and FDR values for each gene, together with their level of significance in the Lawrence study. Genes highlighted in red and orange are those shown to be highly significantly mutated (FDR≤0.001) and significantly mutated (FDR≤0.1) respectively in the Lawrence study. Genes highlighted in blue are potential candidate cancer genes shown to reach significance (FDR≤0.1) in the PAML analysis, however do not reach significance in the Lawrence study. Genes highlighted in purple are examples of genes that do not reach significance in the Lawrence study, and have conflicting results in the PAML analysis (with a significant p-value supporting positive selection, but an omega value indicative of negative selection). *R code used to generate plot in Supplementary Appendix D.*

TABLE 5.33: **Ranked list of significant PAML genes in MEL.** The genes found to be significantly mutated in MEL patients in the PAML analysis have been sorted by p-value in ascending order (descending order of significance) in this table. Genes highlighted in red and orange are found to be highly significantly mutated and significantly mutated respectively in the Lawrence study. Table has been truncated to n=35 rows from a total of 113 significant genes for MEL. *R and Perl code used to produce list in Supplementary Appendix E. Full version of table can be found in Supplementary Appendix F.*

	PAML results		Lawrence et al. [2014] p-values			
Gene	Omega	P-value	CV	CL	FN	Combined
NRAS	121.02	2.36e-39	4.11E-13	9.89E-08	4.17E-04	1.11E-16
BRAF	1.62	7.32e-14	2.90E-16	9.89E-08	9.89E-08	1.11E-16
THSD7B	1.69	4.81e-11	NA	NA	NA	NA
SUCO	2.17	2.76e-10	NA	NA	NA	NA
MYOCD	0.96	3.11e-08	7.80E-01	1.03E-05	1.28E-01	9.06E-05
CNTN5	1.57	3.13e-08	NA	NA	NA	NA
ISX	1.13	3.13e-08	NA	NA	NA	NA
CCDC148	4.06	4.84e-08	NA	NA	NA	NA
PPP6C	999.00	1.35e-07	2.25E-07	1.90E-02	9.03E-01	2.75E-06
UNC13C	2.09	1.40e-07	NA	NA	NA	NA
BRPF1	999.00	3.83e-07	NA	NA	NA	NA
GRID2	1.45	3.92e-07	NA	NA	NA	NA
PRB3	999.00	5.02e-07	NA	NA	NA	NA
MCTP1	3.29	9.38e-07	NA	NA	NA	NA
TP53	999.00	1.27e-06	3.54E-15	7.95E-03	7.56E-04	2.22E-16
OR4A16	3.06	2.59e-06	3.58E-02	1.78E-04	2.46E-01	8.37E-05
CDH6	1.60	3.96e-06	NA	NA	NA	NA
ZNF813	3.24	5.97e-06	NA	NA	NA	NA
FAM83B	1.30	8.38e-06	NA	NA	NA	NA
FRMD4B	2.73	1.11e-05	NA	NA	NA	NA
KRTAP5-10	999.00	1.22e-05	NA	NA	NA	NA
C7	1.25	1.42e-05	NA	NA	NA	NA
WDR33	1.27	1.55e-05	NA	NA	NA	NA
VWA3B	1.13	1.64e-05	NA	NA	NA	NA
CDK4	999.00	1.83e-05	5.24E-04	1.26E-01	6.18E-03	2.31E-05
OR52N1	2.27	2.19e-05	1.73E-02	2.54E-04	2.64E-01	6.75E-05
SYT1	1.81	2.31e-05	NA	NA	NA	NA
OR4K5	2.86	2.47e-05	NA	NA	NA	NA
ADAMDEC1	2.22	2.63e-05	NA	NA	NA	NA
PRB4	999.00	3.53e-05	NA	NA	NA	NA
RP1	1.73	4.00e-05	NA	NA	NA	NA
BAAT	999.00	4.41e-05	NA	NA	NA	NA
KLHL13	1.55	5.06e-05	NA	NA	NA	NA
RYR2	1.09	5.68e-05	NA	NA	NA	NA
SLC39A12	3.29	5.86e-05	NA	NA	NA	NA

TABLE 5.34: **Ranked list of recurrent mutations in MEL.** Ranked list of the top 35 most recurrent SNVs in the MEL subset of the Lawrence dataset, sorted by recurrence in descending order. For each unique mutation, the gene, chromosome, position, ref and alt alleles and how many patients the mutation occurs in (recurrence) have been tabulated.

Gene	Mutation				Recurrence
	Chromosome	Position	Ref	Alt	
BRAF	7	140453136	A	T	62
NRAS	1	115256529	T	C	13
KIAA0907	1	155904250	C	T	11
NRAS	1	115256530	G	T	7
C7orf36	7	39605969	G	A	7
UTP11L	1	38478324	G	A	6
SLC30A6	2	32390905	C	T	6
SLC30A6	2	32390904	C	T	6
MRPS31	13	41345346	C	T	6
DSG4	18	28980822	G	A	6
UMPS	3	124449234	G	A	5
RPL37	5	40835322	G	A	5
RBM22	5	150080667	C	T	5
RAC1	7	6426892	C	T	5
PCDH15	10	55583112	G	A	5
MCTP1	5	94289029	C	T	5
LRRC29	16	67260962	C	T	5
ISX	22	35478537	C	T	5
GRID2	4	94344033	G	A	5
C1orf9	1	172501872	G	A	5
ZNF716	7	57528531	G	A	4
TRHDE	12	73046174	C	T	4
TRHDE	12	73015380	C	T	4
THSD7B	2	138373853	G	A	4
STK19	6	31940123	G	A	4
RPS27	1	153963240	T	C	4
PPP6C	9	127912080	G	A	4
NDUFB9	8	125551345	C	T	4
NDUFB9	8	125551344	C	T	4
MYOCD	17	12661487	C	T	4
MYH1	17	10400301	C	T	4
MCTP2	15	94983621	C	T	4
LUZP2	11	25098972	C	T	4
KCNH7	2	163374667	C	T	4
KCNB2	8	73850273	C	T	4

TABLE 5.35: **Cancer gene detection success in melanoma.** Known cancer genes that have been detected as significantly mutated in all three of the aforementioned analyses, overlapping the MutSig analysis in the Lawrence et al. [2014] study, the PAML analysis and the recurrence analysis.

Known cancer gene
BRAF
NRAS
TP53
PPP6C
CDK4

Table 5.35 shows the known cancer genes that have been successfully detected in melanoma by both the PAML and recurrence analyses in this project, and also by the Lawrence et al. [2014] MutSig analysis.

5.2.12 Multiple myeloma (MM)

In the PAML analysis of multiple myeloma, for which 205 patients were analysed, six genes were found to be significantly mutated with q≤0.1 and undergoing positive selection with an omega ratio >1 (Figure 5.17).

Zinc finger protein 717 (ZNF717) was found to be significantly mutated in the PAML analysis, however it was not identified by the MutSig approach. This gene is a zinc finger protein which has not been implicated as having a role in cancer. However, another zinc finger protein, ZNF311, has been suggested as a novel putative tumour suppressor gene, suppressing the growth and invasiveness of gastric cancer [Yu et al., 2013]. Another zinc finger protein, ZNF143, was suggested to be involved in cellular motility through a ZEB1-E-cadherin-linked pathway in colon cancer cells, increasing cell migration and invasion [Paek et al., 2014]. Additionally, ZNF278 was presented as a potential oncogene in colorectal cancer in Tian et al. [2008], shown to promote cell growth, with knockdown of this gene suppressing proliferation. These findings suggest a role for other zinc finger proteins such as ZNF717 as tumour suppressors in gastric cancers, however it is possible that zinc finger proteins have tumourigenic roles in other tumour types. For example, ZNF652 was shown to interact with the breast tumour suppressor CBFA2T3 to repress transcription indicating a role for ZNF652 in breast cancer [Kumar et al., 2006].

Overall just 13 genes were found to be significant in at least one of the analyses (Table 5.36).

83% (5/6) of the significantly mutated PAML genes for MM are hit by recurrent mutation. The top three most significantly mutated MM genes in PAML analysis are NRAS, KRAS and BRAF. These genes also contain the most recurrent mutations (Table 5.37).

Table 5.38 shows the known cancer genes that have been successfully detected in multiple myeloma by both the PAML and recurrence analyses in this project, and also by the Lawrence et al. [2014] MutSig analysis.

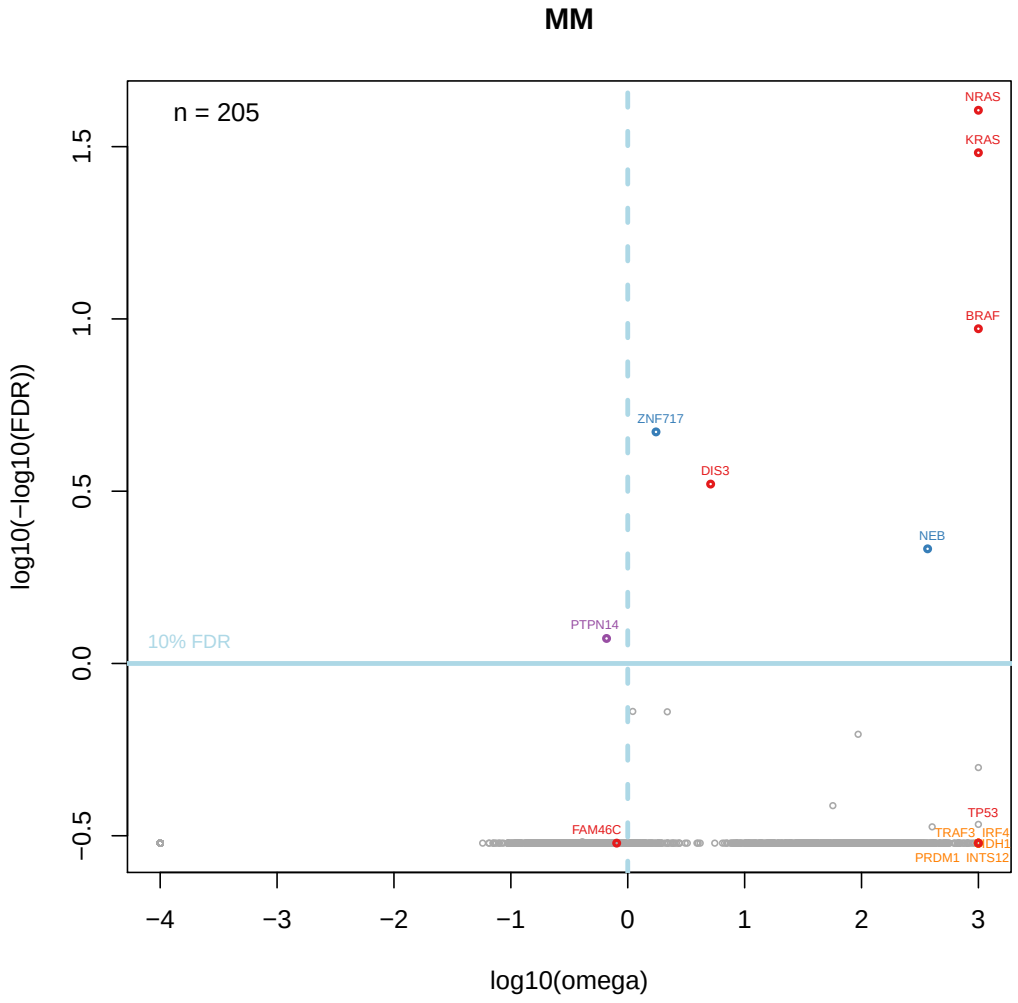


FIGURE 5.17: **Gene-based omega analysis in MM.** Gene-based PAML results have been displayed in this omega plot for 205 MM patients to show the omega ratios and FDR values for each gene, together with their level of significance in the Lawrence study. Genes highlighted in red and orange are those shown to be highly significantly mutated (FDR≤0.001) and significantly mutated (FDR≤0.1) respectively in the Lawrence study. Genes highlighted in blue are potential candidate cancer genes shown to reach significance (FDR≤0.1) in the PAML analysis, however do not reach significance in the Lawrence study. Genes highlighted in purple are examples of genes that do not reach significance in the Lawrence study, and have conflicting results in the PAML analysis (with a significant p-value supporting positive selection, but an omega value indicative of negative selection). *R code used to generate plot in Supplementary Appendix D.*

TABLE 5.36: **Ranked list of significant PAML genes in MM.** The genes found to be significantly mutated in MM patients in the PAML analysis have been sorted by p-value in ascending order (descending order of significance) in this table. Genes highlighted in red and orange are found to be highly significantly mutated and significantly mutated respectively in the Lawrence study. Genes shown below the blue horizontal line are not found to be significant in the PAML analysis but have been tabulated due to their significance in the Lawrence study. *R and Perl code used to produce list in Supplementary Appendix E.*

Gene	PAML results		Lawrence et al. [2014] p-values			
	Omega	P-value	CV	CL	FN	Combined
NRAS	999.00	2.23e-44	6.71E-16	9.99E-08	6.11E-03	1.11E-16
KRAS	999.00	4.07e-34	1.00E-16	9.99E-08	5.99E-01	1.11E-16
BRAF	999.00	6.10e-13	8.88E-05	9.99E-08	2.83E-04	1.98E-10
ZNF717	1.74	3.77e-08	NA	NA	NA	NA
DIS3	5.12	1.14e-06	6.47E-09	2.63E-03	2.08E-02	6.93E-11
NEB	367.60	2.00e-05	NA	NA	NA	NA
TP53	999.00	6.24e-03	1.00E-16	1	1.98E-04	1.11E-16
PRDM1	999.00	8.55e-03	1.82E-03	2.85E-02	1.73E-03	1.26E-05
TRAF3	999.00	6.08e-02	8.05E-07	1	1.57E-01	8.81E-06
FAM46C	0.80	2.11e-01	1.73E-07	2.99E-01	7.64E-03	4.85E-08
INTS12	999.00	2.48e-01	1.25E-04	1.95E-02	4.02E-02	3.79E-06
IRF4	999.00	3.13e-01	2.18E-02	2.91E-04	3.23E-03	2.28E-05
IDH1	999.00	4.36e-01	4.91E-03	1.00E-03	8.19E-01	4.48E-05

TABLE 5.37: **Ranked list of recurrent mutations in MM.** Ranked list of the top 35 most recurrent SNVs in the MM subset of the Lawrence dataset, sorted by recurrence in descending order. For each unique mutation, the gene, chromosome, position, ref and alt alleles and how many patients the mutation occurs in (recurrence) have been tabulated.

Gene	Mutation				Recurrence
	Chromosome	Position	Ref	Alt	
NRAS	1	115256529	T	C	11
NRAS	1	115256530	G	T	9
KRAS	12	25398281	C	T	9
KRAS	12	25380275	T	G	8
PARK2	6	162242837	G	A	5
BRAF	7	140453136	A	T	5
ZNF717	3	75787829	C	G	4
TRAF3IP1	2	239270793	G	T	4
NRAS	1	115258745	C	G	4
NRAS	1	115256528	T	G	4
KCNK1	1	233765476	C	A	4
CACNA2D3	3	54997793	G	T	4
ZNF717	3	75786993	T	A	3
WASH3P	15	102515470	T	C	3
TMEM131	2	98611675	G	T	3
TMEM100	17	53798740	G	T	3
STK3	8	99849608	G	A	3
RHBDL2	1	39400050	C	A	3
RASGEF1C	5	179614811	C	A	3
PXDNL	8	52269072	G	A	3
PTPRK	6	128334135	C	A	3
PPP1R1C	2	182934001	C	T	3
PIK3R4	3	130398753	C	T	3
PDGFRA	4	54248127	G	A	3
NOTCH2	1	120574304	G	A	3
NGEF	2	233877708	G	T	3
MPDZ	9	13224823	C	A	3
LRRN2	1	204623728	G	T	3
LMF1	16	907531	A	T	3
KRAS	12	25398284	C	T	3
KRAS	12	25398284	C	G	3
KRAS	12	25380276	T	C	3
KRAS	12	25378562	C	T	3
IRF4	6	394972	A	G	3
HDAC5	17	42163067	A	G	3

TABLE 5.38: **Cancer gene detection success in multiple myeloma.** Known cancer genes that have been detected as significantly mutated in all three of the aforementioned analyses, overlapping the MutSig analysis in the Lawrence et al. [2014] study, the PAML analysis and the recurrence analysis.

Known cancer gene
KRAS
NRAS
DIS3
BRAF

5.2.13 Ovarian (OV)

In the PAML analysis of ovarian cancer, for which 316 patients were analysed, just two genes were found to be significantly mutated with q≤0.1 and undergoing positive selection with an omega ratio >1 (Figure 5.18).

The only gene identified as significantly mutated in my approach that was not also identified by the MutSig analysis was myosin 3A (MYO3A). This is not a known cancer gene, and hence could potentially be a putative cancer gene. MYO3A is a class III myosin, which contain a conserved N-terminal kinase domain and central motor domain, and a C-terminal tail that varies in sequence between isoforms [Quintero et al., 2013]. There are two isoforms of class III myosins expressed in vertebrates: MYO3A and MYO3B. Both isoforms are localized to actin bundle-based structures in sensory cells, including the calycal process of photoreceptors and the stereocilia of inner ear hair cells, and play a role in maintaining the length of the actin-bundled structures. Disruption of the MYO3A gene is associated with nonsyndromic deafness. However, studies have found that class III myosins are not only expressed in the inner ear and retina but also in the brain, intestine, and testes. Furthermore, modifications in the MYO3A gene were found to be associated with bladder cancer in which MYO3A was found along with five other genes to be a gene methylation biomarker for this cancer type [Chung et al., 2011], and colon cancer in which a genome-wide association study found that single nucleotide polymorphisms in the intronic region of the MYO3A gene were associated with an increased risk for this cancer type [Lascorz et al., 2010].

Overall only six genes were found to be significant across both analyses (Table 5.39).

Of the significant genes from PAML analysis, 100% (2/2) were found to be hit by recurrent mutation, including TP53 which is the most significantly mutated gene in both the PAML and Lawrence analyses. This gene is also hit by multiple recurrent mutations (Table 5.40).

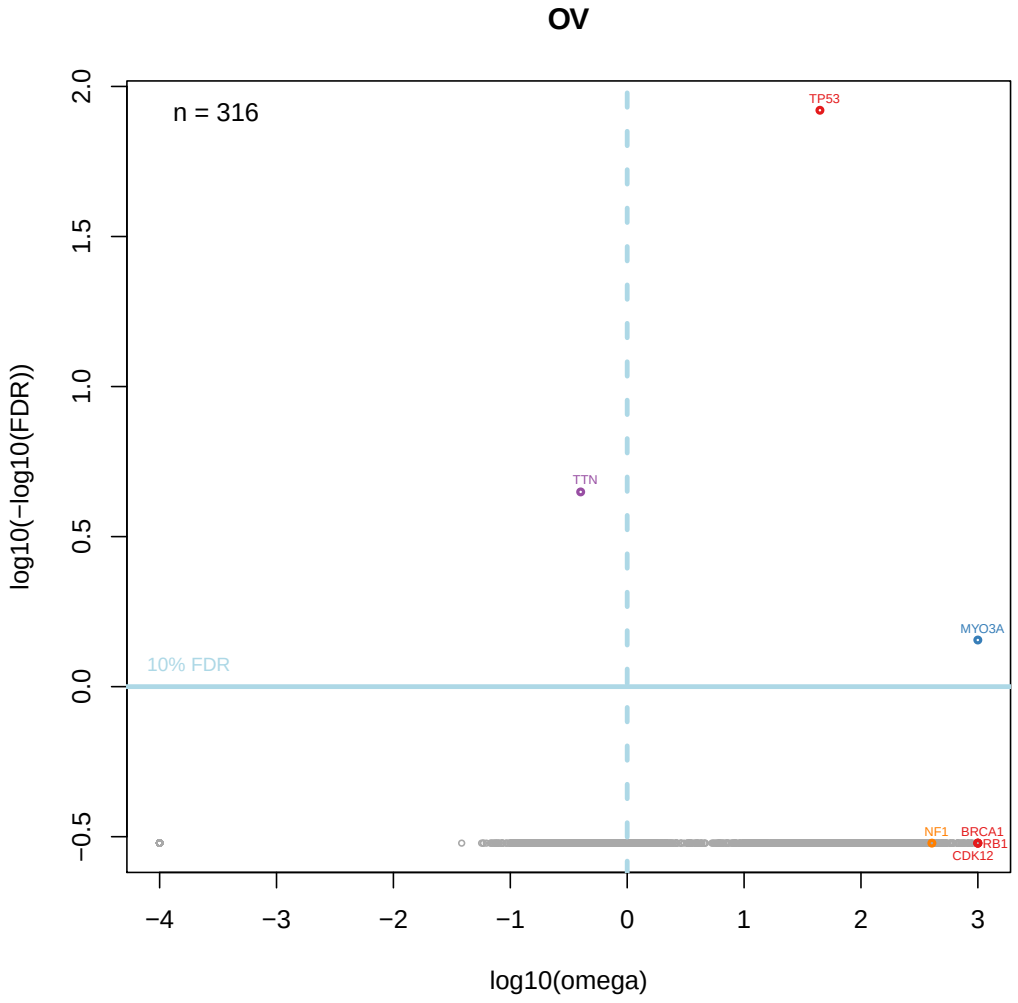


FIGURE 5.18: **Gene-based omega analysis in OV.** Gene-based PAML results have been displayed in this omega plot for 316 OV patients to show the omega ratios and FDR values for each gene, together with their level of significance in the Lawrence study. Genes highlighted in red and orange are those shown to be highly significantly mutated (FDR≤0.001) and significantly mutated (FDR≤0.1) respectively in the Lawrence study. Genes highlighted in blue are potential candidate cancer genes shown to reach significance (FDR≤0.1) in the PAML analysis, however do not reach significance in the Lawrence study. Genes highlighted in purple are examples of genes that do not reach significance in the Lawrence study, and have conflicting results in the PAML analysis (with a significant p-value supporting positive selection, but an omega value indicative of negative selection). *R code used to generate plot in Supplementary Appendix D.*

TABLE 5.39: **Ranked list of significant PAML genes in OV.** The genes found to be significantly mutated in OV patients in the PAML analysis have been sorted by p-value in ascending order (descending order of significance) in this table. Genes highlighted in red and orange are found to be highly significantly mutated and significantly mutated respectively in the Lawrence study. Genes shown below the blue horizontal line are not found to be significant in the PAML analysis but have been tabulated due to their significance in the Lawrence study. *R and Perl code used to produce list in Supplementary Appendix E.*

Gene	PAML results		Lawrence et al. [2014] p-values			
	Omega	P-value	CV	CL	FN	Combined
TP53	44.55	1.20e-87	3.90E-15	5.36E-08	5.36E-08	1.11E-16
MYO3A	999.00	2.67e-05	NA	NA	NA	NA
CDK12	999.00	2.45e-02	9.21E-09	1	2.92E-01	1.42E-07
NF1	404.17	2.98e-02	7.17E-08	1	1.19E-01	9.58E-07
RB1	999.00	1.09e-01	2.86E-09	1	5.95E-01	4.74E-08
BRCA1	999.00	1.66e-01	2.38E-12	1	1.62E-01	5.64E-11

Table 5.41 shows the known cancer genes that have been successfully detected in ovarian cancer by both the PAML and recurrence analyses in this project, and also by the Lawrence et al. [2014] MutSig analysis.

TABLE 5.40: **Ranked list of recurrent mutations in OV.** Ranked list of the top 35 most recurrent SNVs in the OV subset of the Lawrence dataset, sorted by recurrence in descending order. For each unique mutation, the gene, chromosome, position, ref and alt alleles and how many patients the mutation occurs in (recurrence) have been tabulated.

Gene	Mutation				Recurrence
	Chromosome	Position	Ref	Alt	
TP53	17	7578190	T	C	11
TP53	17	7577120	C	T	11
TP53	17	7577538	C	T	9
TP53	17	7578406	C	T	8
TP53	17	7578265	A	G	7
TP53	17	7577121	G	A	7
TP53	17	7577539	G	A	6
TP53	17	7578461	C	A	5
TP53	17	7578403	C	T	5
TP53	17	7577094	G	A	5
TP53	17	7578394	T	C	4
TP53	17	7577559	G	A	4
TP53	17	7578275	G	A	3
TP53	17	7578271	T	C	3
TP53	17	7578263	G	A	3
TP53	17	7577548	C	T	3
TP53	17	7577022	G	A	3
TP53	17	7574003	G	A	3
ZNF536	19	31039137	G	C	2
WNT11	11	75898143	C	T	2
UXS1	2	106761805	C	G	2
UBTF	17	42292477	A	G	2
TUBA3C	13	19751364	C	T	2
TRPV6	7	142583147	G	T	2
TRPC7	5	135692447	T	A	2
TP53	17	7579311	C	A	2
TP53	17	7578556	T	C	2
TP53	17	7578555	C	T	2
TP53	17	7578555	C	A	2
TP53	17	7578534	C	A	2
TP53	17	7578454	G	A	2
TP53	17	7578442	T	C	2
TP53	17	7578370	C	A	2
TP53	17	7578290	C	T	2
TP53	17	7578268	A	C	2

TABLE 5.41: **Cancer gene detection success in ovarian cancer.** Known cancer genes that have been detected as significantly mutated in all three of the aforementioned analyses, overlapping the MutSig analysis in the Lawrence et al. [2014] study, the PAML analysis and the recurrence analysis.

Known cancer gene
TP53

5.3 Discussion

Evolutionary analysis in PAML was performed on mutations from the published dataset of Lawrence et al. [2014] to identify genes containing driver mutations in 13 different types of cancer, and the results were compared to the published results of Lawrence et al. [2014].

As well as showing that both methods can detect known cancer genes, a novel candidate driver gene was identified in colorectal cancer from the PAML analysis.

MutSig is a less well-established framework in comparison to PAML, with several underlying assumptions. However, Lawrence et al. [2014] is still a complementary approach that has been used in combination with the PAML method.

5.3.1 Comparison of cancer gene detection methods

Clearly both methods pick up many of the same genes, however PAML misses some of the genes that MutSig picks up and vice versa. Some of the difference can be attributed to the fact that in the PAML analysis only the coding mutations have been used whereas Lawrence have used both coding and non-coding. Also before PAML analysis INDELs and larger substitutions were filtered out to leave just SNVs since the codon model used in PAML is not suitable for larger mutations.

However it is not just the filtering of the data that differs but the framework of the tests. The statistical test performed in PAML is quite different to that of MutSig.

5.3.1.1 Mutation clustering in genes

MutSig may be superior at finding genes with clustered mutations in cancer. For example in CRC, the gene BRAF is detected using MutSig when it is not in PAML. This could be partly accounted for by one of the statistical tests that MutSig uses,

MutSigCL, to account for clustering, which appears to be occurring in this gene. PAML may not be as powerful in these cases.

5.3.1.2 Technical difficulties in PAML

Some of the genes that Lawrence have found to be significant, PAML has failed to return results for. Probably issues are:

1. There are no mutations or only one point mutation per gene, which means there will be no alignment to calculate model parameters from. In these cases, Lawrence may have significant results from INDELs.
2. It is rare but occasionally, particularly for short compositionally biased genes, PAML is unable to parametise its models and fails its convergence test.

5.3.1.3 Power to detect different modes of selection in PAML

Lawrence et al. [2014] did not have the power to detect both positive and negative selection occurring at the same time since they have not used selection to measure, whereas the advantage of using the PAML approach was that is was a well validated approach and may be able to pick up signals of genes undergoing both positive and purifying selection.

5.3.1.4 Higher false-positive rate in PAML

PAML seems to be finding very large genes (e.g. TTN) significant, although these are suspected to be false-positives since they often come up in such studies as false-positives. Lawrence do not get these same spurious results, which suggests that their methodology is superior at removing false-positives [Lawrence et al., 2013].

5.3.2 Novel candidate cancer gene in colon adenocarcinaom: DNMT1

The most interesting discovery by PAML is the detection of DNMT1 in colorectal cancer as a significantly mutated gene. This gene was not detected in the Lawrence study, and is therefore a good candidate cancer gene for experimental validation.

5.3.3 Relating tissue of origin and mutation to path of selection

The aim of this chapter was to investigate how the tissue of origin relates to the genes hit by mutation. It can be seen from the above results in PAML and both the Lawrence study that there is merit in partitioning data by tissue of origin, since different genes are hit by driver mutations in different cancer types, showing that different mechanisms are occurring in different organs, as is already known to be the case.

5.3.4 Complex codon model vs recurrent mutation count

Many of the genes found to be significantly mutated using the evolutionary method in PAML are also often the genes with the highest recurrence of a particular mutation, often containing more than one particular recurrent mutation. However many genes identified through PAML were not recurrently mutated, therefore it is necessary to use such a complex codon model to detect cancer genes rather than just counting recurrent mutations in a gene. Also the recurrent counts stated in this chapter are inclusive of both synonymous and non-synonymous mutations. The number of synonymous mutations have not been stated, but these could be involved in processes such as alternate splicing. Most are expected to be within CpG dinucleotides, which they are, however this has not been shown in this chapter.

5.4 Methods

Lawrence cancer-specific mutations for each gene were downloaded in MAF format, for all 4728 patients over all 21 cancer types. INDELs and substitutions longer than one nucleotide were removed, along with non-coding SNVs leaving 4712 patients for evolutionary analysis.

5.4.1 Partitioning data by tissue of origin

Patients were partitioned by cancer type prior to editing the cancer-specific mutations onto the reference transcripts for each gene. From the available 21 cancer types, only 13 were chosen for analysis in PAML. The criteria for this being that they had sufficient patient numbers available for increased power in PAML, so only cancer types with at least 150 patients were used, with the exception of melanoma (MEL) which only had 118 patients. Melanoma was included because it is known to have such a interesting and specific mutational profile caused by exposure to UV radiation, as well as a generally high mutation rate (mutator phenotype).

5.4.2 PAML analysis

For each cancer type, patient alignments edited with the Lawrence cancer-specific mutations were run through PAML in the same way as was done for the whole TCGA dataset, on a per gene basis as before.

An omega plot was generated in R for each of the 13 cancer types in the same way as has been done for the whole-dataset TCGA analysis in Chapter 4. Tables were also created listing the most significant genes for each cancer type.

5.4.3 Recurrent mutations

A basic analysis to find the recurrence of particular mutations across all patients was carried out on each of the 13 cancer types, by simply counting how many patients carried the mutation. The top 35 most recurrent mutations (mutations occurring most frequently in the dataset) were tabulated for each cancer type. This analysis included both non-synonymous and synonymous mutations, over all coding and non-coding SNVs.

Chapter 6

Evolutionary sub-type analysis: stratification by mutation spectra

6.1 Introduction

Mutation spectra does not necessarily cluster by cancer type as has been discovered in Chapter 3 over both the TCGA dataset and the Lawrence dataset. This was the motivation for the analysis in this chapter, in which different types of mutation spectra have been accounted for in order to further inform our evolutionary analysis.

Addressing the question of how mutation spectra influences the genes hit by driver mutation in cancer, the published Lawrence data used in Chapter 5 has been stratified by mutation spectra before evolutionary analysis in PAML. Results from the different types of mutation spectra were compared to find if patterns of positive selection occur in the same places in these different groups of cancer, or if as hypothesised the patterns differ by mutation spectra.

As in Chapter 5, only the coding cancer-specific SNVs have been used for this analysis.

6.2 Results

The data was split into six discrete groups based on the mutational profiles of the patients.

Again to interpret the results from PAML plots akin to the ones generated in Chapter 4 and Chapter 5 have been plotted for each of the six different mutation spectra groups.

Gene lists were also produced for each group containing only the genes with significant q-values (FDR) in ranked order of significance. Genes with an omega <1 have not been included.

6.2.1 Mutational signatures across Lawrence dataset

The 4,728 patients in the Lawrence dataset were split into six groups each with a distinct mutational signature, based on the relative proportions of their single nucleotide changes (of which there are six classes). This was done using cluster analysis in R to group together patients with similar mutation spectra (Figure 6.1). The six groups were chosen arbitrarily, using a height cut-off of 0.65 to ensure six substantially sized groups of patients. However, 138 of the 4,728 patients had to be excluded from subsequent analysis since they were not present in any of the large clustered groups. The excluded patients fell into 12 small distinct groups at a height cut-off of 0.65. This indicated that the mutation spectra of these excluded patients were varied and not common across the dataset, and evolutionary analysis of these 12 small sample groups would not have much power in detecting cancer genes. 4,590 patients were split across the six groups chosen for analysis: 453 in group 1, 1,073 in group 2, 375 in group 3, 457 in group 4, 1,526 in group 5 and 706 in group 6.

The proportions of each of the six classes of single nucleotide change in each mutational signature is shown in Figure 6.2.

The specific spectra of Signature 2 exhibits an elevated proportion of C→T transition mutations compared to any other mutation. Signatures 5 and 6 show a similar pattern,

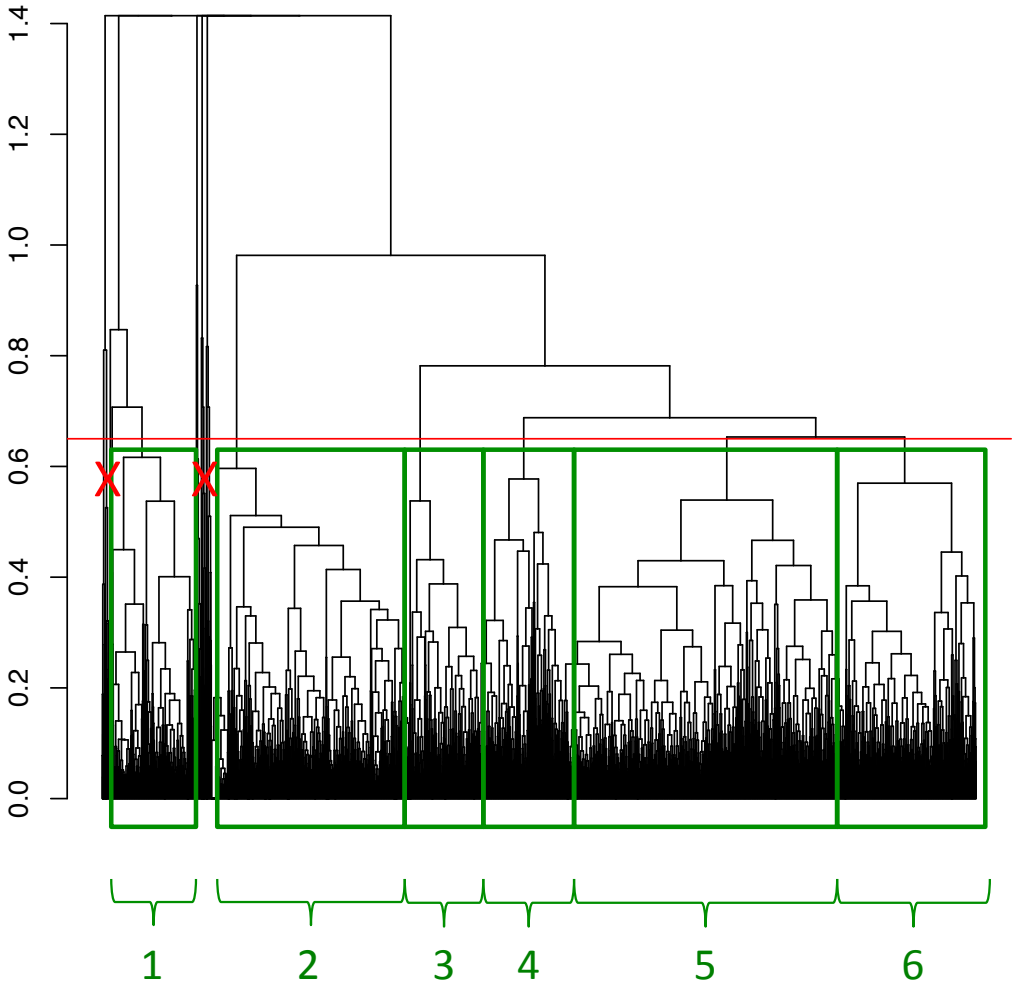
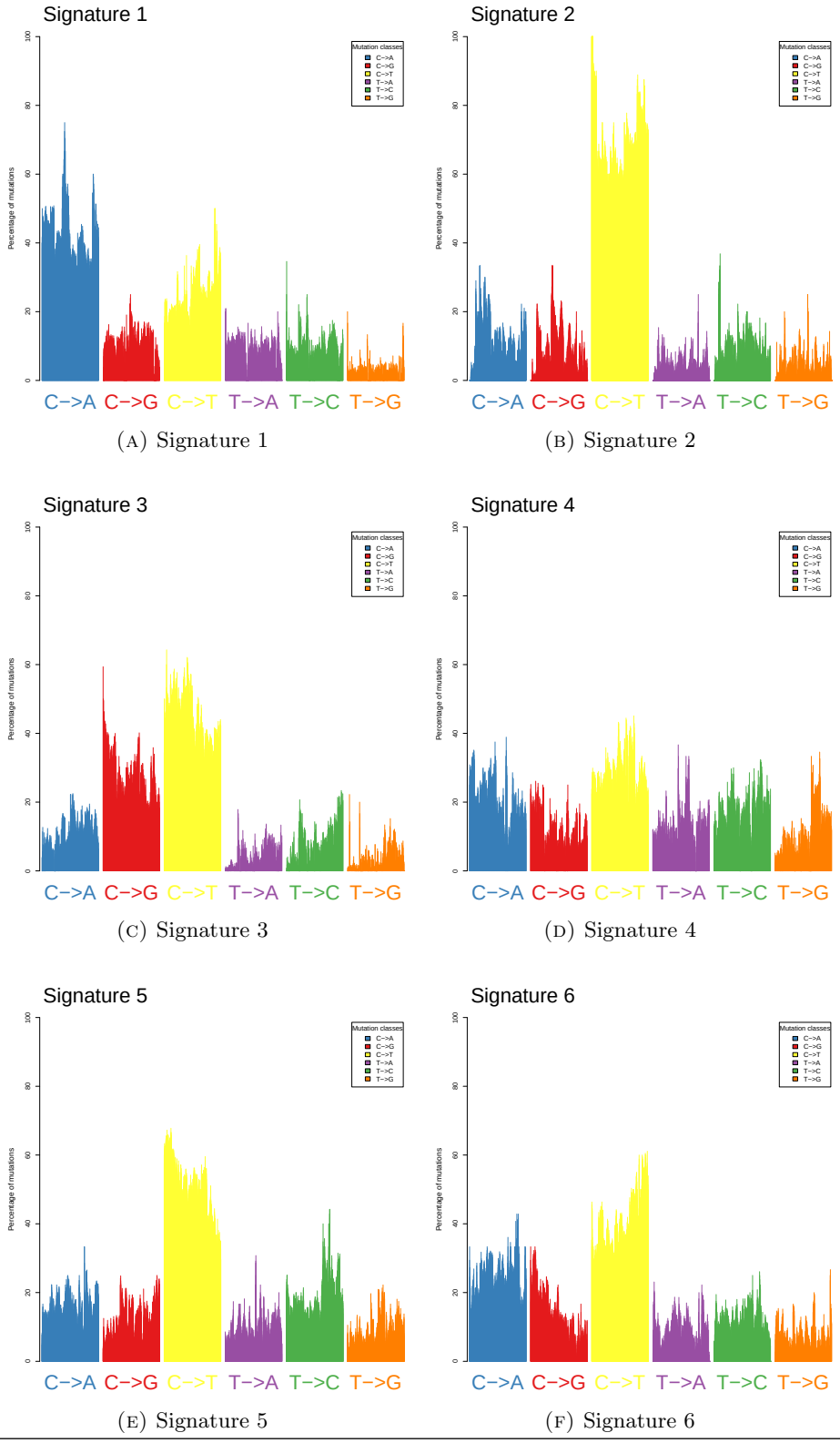


FIGURE 6.1: **Lawrence patients clustered by mutation spectra.** Dendrogram to show how the 4,728 patients in the Lawrence dataset have been clustered according to their relative proportions of each of the six classes of single nucleotide changes. The red horizontal line shows the arbitrary height (h=0.65) at which the dataset has been split into 18 clustered groups (clusters with a distance of less than 0.65 between them). Green rectangles highlight the six largest distinct mutation spectra signatures (including 4,590 patients) within these 18 groups, which were used for subsequent analysis. 12 small groups containing 138 patients altogether were excluded from analysis, due to their small group sizes (marked with red X). The y-axis represents the height, which is a measure of the similarity of the clusters (greater the height, greater the difference). The patients are plotted along the x-axis in an order according to their relationship of similarity with other patients. Along the x-axis: 1 = 'Signature 1'; 2 = 'Signature 2'; 3 = 'Signature 3'; 4 = 'Signature 4'; 5 = 'Signature 5'; 6 = 'Signature 6'.



(A) Signature 1

(B) Signature 2

(C) Signature 3

(D) Signature 4

(E) Signature 5

(F) Signature 6

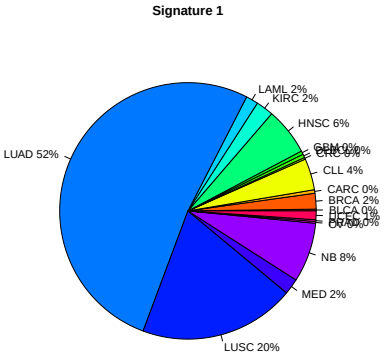
FIGURE 6.2: **Mutational signatures across the Lawrence dataset.** The relative proportions of each of the six classes of single nucleotide changes over the 4,728 Lawrence patients in (A) Signature 1, (B) Signature 2, (C) Signature 3, (D) Signature 4, (E) Signature 5 and (F) Signature 6. Patients have been plotted along the x-axis, in the same order in each plot for comparison. *R code used to generate plots in Supplementary Appendix G.*

also exhibiting a higher frequency of C→T mutations, however it is not as pronounced as in Signature 2. Signature 4 has a more uniform mutation profile, with a roughly even distribution of each of the six mutation classes. However there is generally an elevated proportion of this particular mutation class observed in the human population. This is due to instability at CpG islands and deamination. It is this elevated rate of C→T that is thought to lead to a higher rate of transitions over transversions in the genome.

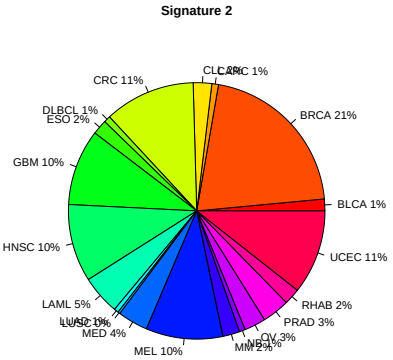
Figure 6.3 shows how the six mutational signatures are enriched for certain tumour types. A statistical test for enrichment of tumour types within the six different signatures was performed in R using Pearson's Chi-squared test for count data. Table 6.1 shows the observed number of patients within each of the six groups for each cancer type, as well as the results from the chi-squared test: the expected number of patients per group that would be expected under the null hypothesis for each tumour type (no significant difference between groups); and the p-value to test the significance of the statistical hypothesis. It can be seen that all p-values for all cancer types are highly significant (at the 1% significance level, with p-value <0.01), rejecting the null hypothesis for each test and indicating that for all of the 21 different cancer types the observed patient counts are not evenly spread amongst the six signatures and that there is enrichment of tumour type in certain signatures. The most significant p-value was calculated for LUAD (p-value = 9.19e-115), showing that there is significant enrichment of this cancer type in Signature 1 (235 LUAD patients observed compared to 66 expected).

Signature 1 shows the most significant enrichment of tumour type, with 52% of the patients within this signature having lung adenocarcinoma (LUAD) and 20% with lung squamous cell carcinoma (LUSC). This specific signature exhibits an elevated prevalence of C→A substitutions which is consistent with exposure to the polycyclic aromatic hydrocarbons in tobacco smoke [Alexandrov et al., 2013, Lawrence et al., 2013], which is known to cause lung cancer. Therefore this result is expected.

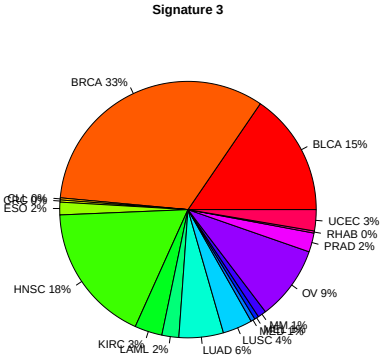
Signature 2 has a very high prevalence of C→T mutations which is seen in melanoma as a result of mutations caused by the misrepair of ultraviolet-induced covalent bonds between adjacent pyrimidines [Lawrence et al., 2013]. However, in Figure 6.3, this



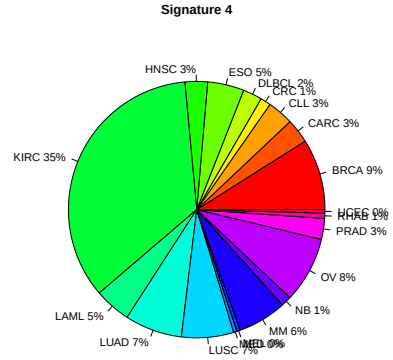
(A) Signature 1



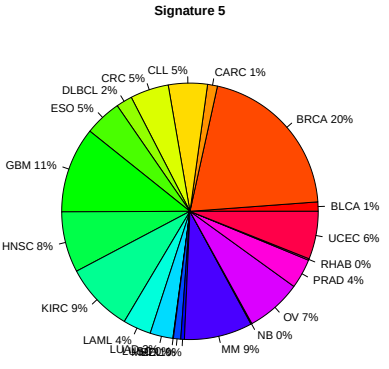
(B) Signature 2



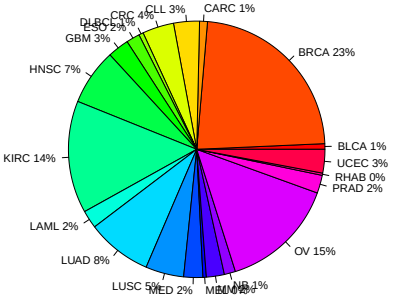
(C) Signature 3



(D) Signature 4



(E) Signature 5



(F) Signature 6

FIGURE 6.3: **Enrichment of tumour types within each mutational signature.** The number of patients with a particular tumour type, out of the total number of patients in each mutational spectra group, has been plotted as a pie chart for (A) Signature 1, (B) Signature 2, (C) Signature 3, (D) Signature 4, (E) Signature 5 and (F) Signature 6. The percentages at which each tumour type is present within the mutational signature are also shown.

TABLE 6.1: **Statistical test for tumour type enrichment within mutational signature groups.** Using Pearson's Chi-squared test for count data, a goodness-of-fit test was performed in R for each tumour type, to test whether the observed counts of patients are statistically significantly different between each of the six mutational spectra groups for a given tumour type. The null hypothesis is that the observed counts are all equal amongst the six groups. For each tumour type, the p-value is statistically significant, rejecting the null hypothesis and showing that the patient counts are are not evenly split between groups. Expected counts are rounded to the nearest integer. P-values are rounded to 2 decimal places.

Disease	Observed counts						Expected counts under null hypothesis	P-value for the test
	1	2	3	4	5	6		
BLCA	1	16	58	0	18	5	16	1.50e-29
BRCA	9	223	124	41	311	162	145	1.06e-92
CARC	2	9	0	14	19	7	9	1.29e-05
CLL	18	25	1	15	75	23	26	6.95e-25
CRC	1	123	1	6	74	28	39	9.21e-67
DLBCL	2	9	0	11	31	4	10	3.35e-13
ESO	0	20	6	21	71	12	22	1.48e-30
GBM	2	103	0	0	165	19	48	1.06e-106
HNSC	27	105	66	13	117	50	63	6.50e-28
KIRC	10	0	13	159	133	100	69	1.69e-74
LAML	7	53	8	21	54	16	27	2.14e-17
LUAD	235	6	21	33	44	58	66	9.19e-115
LUSC	89	3	14	30	1	34	29	2.42e-38
MED	9	41	2	2	14	17	14	1.38e-14
MEL	0	104	2	2	7	3	20	7.20e-92
MM	0	23	4	28	131	16	34	1.61e-74
NB	34	8	0	6	2	10	10	5.75e-15
OV	1	29	35	38	107	103	52	2.18e-36
PRAD	1	37	9	12	57	16	22	5.77e-20
RHAB	0	22	1	3	3	2	5	4.05e-13
UCEC	5	114	10	2	92	21	41	3.09e-62

signature is shown to be enriched in breast cancer (BRCA) present at 21% whereas melanoma (MEL) is present at only 10%. Gastrointestinal tumours such as colorectal cancer also show extremely high frequencies of C→T (at CpG dinucleotides), and in this signature colorectal cancer (CRC) is present at 11%.

Signature 3 has been shown to have elevated rates of both C→G and C→T substitutions. This signature is also shown to be enriched with breast cancer (BRCA) present at 33%, head and neck cancer (HNSC) present at 18% and bladder cancer (BLCA) present at 15%. Head and neck cancer and bladder cancer have previously been shown to exhibit elevated rates of C→G and C→T mutations [Alexandrov et al., 2013, Lawrence et al., 2013], supporting these results. This pattern is characteristic of mutations caused by the APOBEC family of cytidine deaminases, which are innate immunity enzymes that are induced by certain classes of viruses. The human papillomavirus (HPV) has previously been implicated in head and neck cancers, indicating that a subset of both bladder (also suggested by [Lawrence et al., 2013]) and breast cancers could have a viral aetiology.

Signature 4, which exhibits a uniform distribution of the six classes of substitution, is enriched with kidney clear cell cancer (KIRC) with 35% of patients exhibiting this cancer type.

Signatures 5 and 6 show an enrichment of breast cancer (BRCA) occurring at 20% and 23% respectively. Both these signatures exhibit elevated C→T, however signature 6 has a slightly higher prevalence of C→A and C→G mutations than signature 5, and signature 5 with a slightly higher prevalence of T→C than signature 6.

For all six signatures, the sequence context immediately 5' and 3' to the mutated nucleotide was not considered, which would help to further elucidate the mutational processes that may be underlying each specific signature.

6.2.1.1 Signature 1

Signature 1 (Figure 6.2) contains 453 patients. These patients have mostly C→A changes.

Analysis in PAML has revealed that the most significantly mutated genes in these patients are KRAS, TP53 and PIK3CA (FIgure 6.4) and other well known cancer genes have also been detected as containing driver mutations (e.g. PTEN, BRAF and EGFR).

In total, 38 genes were found to be significantly mutated in Signature 1, and the top most significant 35 of these have been tabulated in Table 6.2 with their omega estimate, p-values and gene descriptions.

GO term analysis in GOrilla has produced the top ten processes enriched for the significant genes in this analysis (Table 6.3).

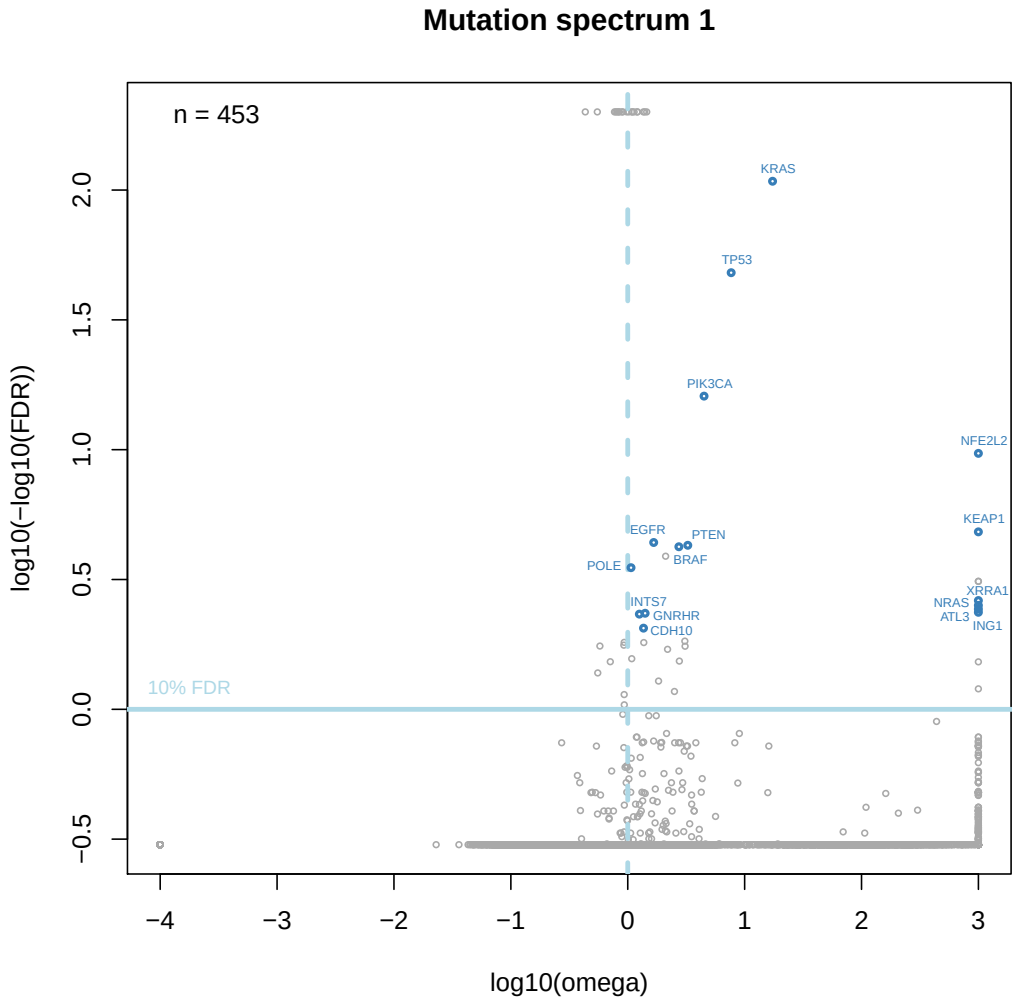


FIGURE 6.4: **Gene-based omega analysis in PAML for Signature 1.** For each gene in the group of 453 patients in Signature 1, the omega estimate and FDR value from PAML analysis has been plotted to show the patterns of selection across the dataset. *R code used to generate plot in Supplementary Appendix H.*

TABLE 6.2: **Ranked list of significantly mutated genes in Signature 1.** Significantly mutated genes (omega>1 and FDR<0.1) from PAML analysis of Signature 1 have been tabulated in ascending order by p-value (descending order of significance). List has been truncated to n=35 rows out of a total of 38 significant genes. *Full table in Supplementary Appendix J. Code used to generate table in Supplementary Appendix I.*

Gene	Omega	P-value	Description
DNAH9	1.39	0.00e+00	dynein, axonemal, heavy chain 9
USH2A	1.09	0.00e+00	Usher syndrome 2A (autosomal recessive, mild)
FLG	1.45	0.00e+00	filaggrin
XIRP2	1.02	0.00e+00	xin actin-binding repeat containing 2
CSMD3	1.20	0.00e+00	CUB and Sushi multiple domains 3
MUC17	1.36	0.00e+00	mucin 17, cell surface associated
PKHD1	1.09	0.00e+00	polycystic kidney and hepatic disease 1 (autosomal recessive)
MUC16	1.21	0.00e+00	mucin 16, cell surface associated
PCLO	1.13	0.00e+00	piccolo presynaptic cytomatrix protein
KRAS	17.33	9.69e-112	Kirsten rat sarcoma viral oncogene homolog
TP53	7.67	1.18e-51	tumor protein p53
PIK3CA	4.50	1.16e-19	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha
NFE2L2	999.00	3.06e-13	nuclear factor, erythroid 2-like 2
KEAP1	999.00	2.28e-08	kelch-like ECH-associated protein 1
EGFR	1.67	6.48e-08	epidermal growth factor receptor
PTEN	3.26	8.63e-08	phosphatase and tensin homolog
BRAF	2.74	1.02e-07	v-raf murine sarcoma viral oncogene homolog B
KMT2D	2.11	2.31e-07	lysine (K)-specific methyltransferase 2D
POLE	1.07	5.68e-07	polymerase (DNA directed), epsilon, catalytic subunit
MIB1	999.00	1.47e-06	mindbomb E3 ubiquitin protein ligase 1
HRAS	999.00	1.53e-06	Harvey rat sarcoma viral oncogene homolog
XRRA1	999.00	4.84e-06	X-ray radiation resistance associated 1
NRAS	999.00	6.44e-06	neuroblastoma RAS viral (v-ras) oncogene homolog
ATL3	999.00	8.00e-06	atlastin GTPase 3
ING1	999.00	9.63e-06	inhibitor of growth family, member 1
GNRHR	1.41	1.04e-05	gonadotropin-releasing hormone receptor
INTS7	1.26	1.11e-05	integrator complex subunit 7
CDH10	1.36	2.14e-05	cadherin 10, type 2 (T2-cadherin)
C10ORF68	3.08	3.66e-05	Homo sapiens coiled-coil domain containing 7 (CCDC7), transcript variant 5, mRNA.
RAD50	1.37	4.06e-05	RAD50 homolog (S. cerevisiae)
SMC4	3.09	4.99e-05	structural maintenance of chromosomes 4
CAPN5	2.19	5.69e-05	calpain 5
XPO4	1.08	7.95e-05	exportin 4
COL11A1	2.76	8.75e-05	collagen, type XI, alpha 1
DROSHA	999.00	9.35e-05	drosha, ribonuclease type III

TABLE 6.3: **Enriched GO terms in Signature 1.** Top 10 enriched GO terms in Signature 1, using process ontology from GOrilla.

Process description	P-value	FDR q-value	Enrichment (N, B, n, b)
fibroblast growth factor receptor signaling pathway	1.24E-9	1.64E-5	23.72 (18208,166,37,8)
neurotrophin TRK receptor signaling pathway	6.64E-8	4.37E-4	14.26 (18208,276,37,8)
neurotrophin signaling pathway	7.42E-8	3.26E-4	14.06 (18208,280,37,8)
Fc-epsilon receptor signaling pathway	7.65E-8	2.52E-4	18.72 (18208,184,37,7)
positive regulation of Rac protein signal transduction	7.7E-8	2.03E-4	295.26 (18208,5,37,3)
epidermal growth factor receptor signaling pathway	1.26E-7	2.77E-4	17.40 (18208,198,37,7)
ERBB signaling pathway	1.4E-7	2.63E-4	17.14 (18208,201,37,7)
activation of MAPKK activity	1.69E-7	2.78E-4	39.06 (18208,63,37,5)
response to radiation	2.18E-7	3.19E-4	9.82 (18208,451,37,9)
Fc receptor signaling pathway	4.92E-7	6.48E-4	14.23 (18208,242,37,7)

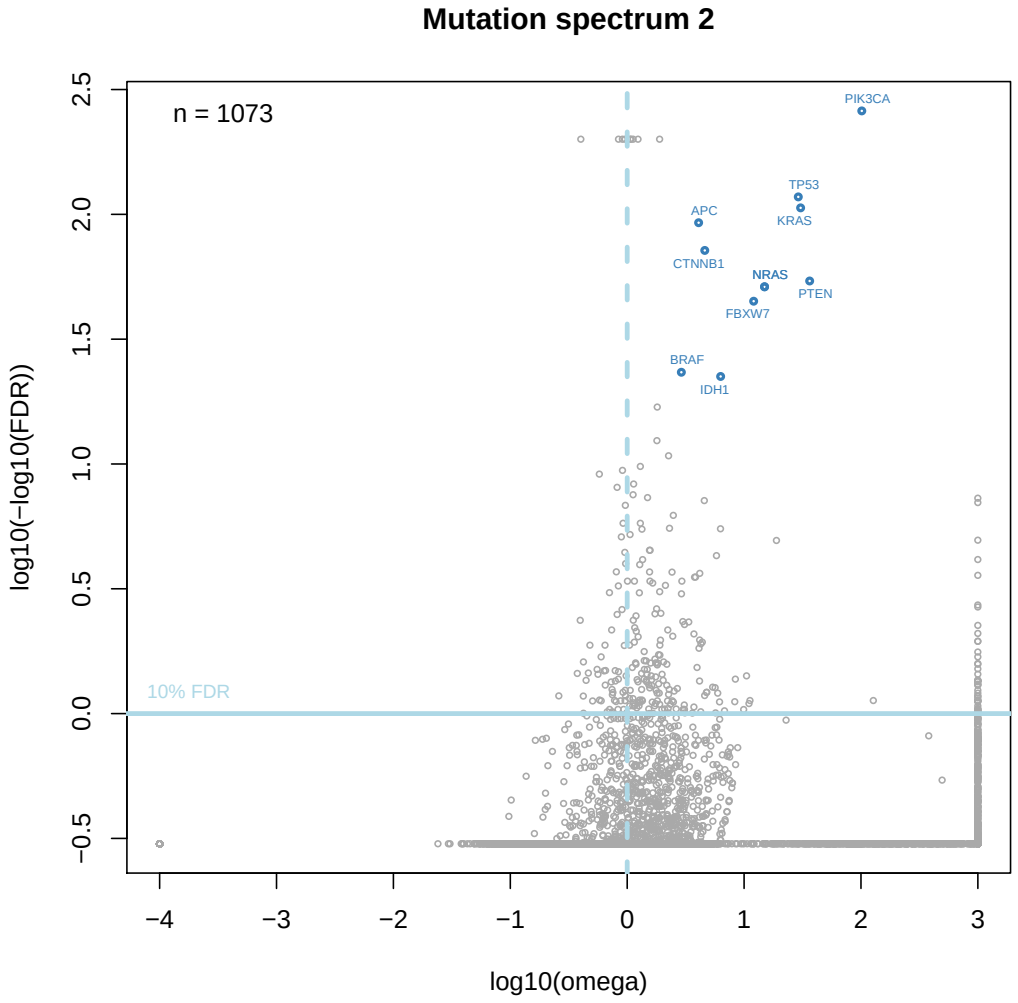


FIGURE 6.5: **Gene-based omega analysis in PAML for Signature 2.** For each gene in the group of 1,073 patients in Signature 2, the omega estimate and FDR value from PAML analysis has been plotted to show the patterns of selection across the dataset. *R code used to generate plot in Supplementary Appendix H.*

6.2.1.2 Signature 2

Signature 2 (Figure 6.2) contains 1073 patients. These patients have mostly C→T changes.

In total, 218 genes were found to be significantly mutated in Signature 2, and the top most significant 35 of these have been tabulated in Table 6.4 with their omega estimate,

p-values and gene descriptions.

GO term analysis in GOrilla has produced the top ten processes enriched for the significant genes in this analysis (Table 6.5).

TABLE 6.4: **Ranked list of significantly mutated genes in Signature 2.** Significantly mutated genes (omega>1 and FDR<0.1) from PAML analysis of Signature 2 have been tabulated in ascending order by p-value (descending order of significance). List has been truncated to n=35 rows out of a total of 218 significant genes. *Full table in Supplementary Appendix J. Code used to generate table in Supplementary Appendix I.*

Gene	Omega	P-value	Description
DNAH5	1.08	0.00e+00	dynein, axonemal, heavy chain 5
FLG	1.89	0.00e+00	filaggrin
LRP1B	1.06	0.00e+00	low density lipoprotein receptor-related protein 1B
MUC16	1.24	0.00e+00	mucin 16, cell surface associated
FAT4	1.12	0.00e+00	FAT atypical cadherin 4
PIK3CA	101.61	1.23e-263	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha
TP53	29.08	2.16e-121	tumor protein p53
KRAS	30.44	4.65e-110	Kirsten rat sarcoma viral oncogene homolog
APC	4.08	2.00e-96	adenomatous polyposis coli
CTNNB1	4.61	1.68e-75	catenin (cadherin-associated protein), beta 1, 88kDa
PTEN	36.39	7.22e-58	phosphatase and tensin homolog
NRAS	14.96	5.13e-55	neuroblastoma RAS viral (v-ras) oncogene homolog
FBXW7	12.07	1.34e-48	F-box and WD repeat domain containing 7, E3 ubiquitin protein ligase
BRAF	2.91	5.25e-27	v-raf murine sarcoma viral oncogene homolog B
IDH1	6.30	4.27e-26	isocitrate dehydrogenase 1 (NADP+), soluble
CDKN2A	1.81	1.47e-20	cyclin-dependent kinase inhibitor 2A
ERBB2	1.80	4.94e-16	v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2
ISX	2.26	2.13e-14	intestine-specific homeobox
EGFR	1.30	2.28e-13	epidermal growth factor receptor
BCOR	1.14	7.48e-12	BCL6 corepressor
MACF1	1.12	4.89e-11	microtubule-actin crosslinking factor 1
IDH2	1.49	7.95e-11	isocitrate dehydrogenase 2 (NADP+), mitochondrial
HRAS	999.00	9.01e-11	Harvey rat sarcoma viral oncogene homolog
XPO1	4.57	1.35e-10	exportin 1 (CRM1 homolog, yeast)
PPP6C	999.00	1.87e-10	protein phosphatase 6, catalytic subunit
SUCO	2.48	1.20e-09	SUN domain containing ossification factor
GRID2	1.30	3.47e-09	glutamate receptor, ionotropic, delta 2
ARID1A	2.30	6.51e-09	AT rich interactive domain 1A (SWI-like)
PIK3R1	6.28	7.09e-09	phosphoinositide-3-kinase, regulatory subunit 1 (alpha)
FGFR2	1.34	7.63e-09	fibroblast growth factor receptor 2
POSTN	1.06	1.44e-08	periostin, osteoblast specific factor
PRAMEF20	999.00	2.78e-08	PRAME family member 20
PRB3	18.92	2.93e-08	proline-rich protein BstNI subfamily 3
RP1	1.57	7.99e-08	retinitis pigmentosa 1 (autosomal dominant)
THSD7B	1.55	8.20e-08	thrombospondin, type I, domain containing 7B

TABLE 6.5: **Enriched GO terms in Signature 2.** Top 10 enriched GO terms in Signature 2, using process ontology from GOrilla.

Process description	P-value	FDR q-value	Enrichment (N, B, n, b)
regulation of cell differentiation	1.45E-6	1.91E-2	2.34 (18208,1376,204,36)
positive regulation of macromolecule metabolic process	2.49E-6	1.64E-2	1.92 (18208,2372,204,51)
fibroblast growth factor receptor signaling pathway	2.77E-6	1.21E-2	5.91 (18208,166,204,11)
neurotrophin TRK receptor signaling pathway	2.93E-6	9.65E-3	4.53 (18208,276,204,14)
neurotrophin signaling pathway	3.46E-6	9.12E-3	4.46 (18208,280,204,14)
regulation of organ morphogenesis	8.43E-6	1.85E-2	5.87 (18208,152,204,10)
positive regulation of gene expression	8.83E-6	1.66E-2	2.19 (18208,1428,204,35)
regulation of cell cycle phase transition	9.18E-6	1.51E-2	4.76 (18208,225,204,12)
regulation of cellular component organisation	9.66E-6	1.41E-2	2.04 (18208,1750,204,40)
positive regulation of Rac protein signal transduction	1.36E-5	1.79E-2	53.55 (18208,5,204,3)

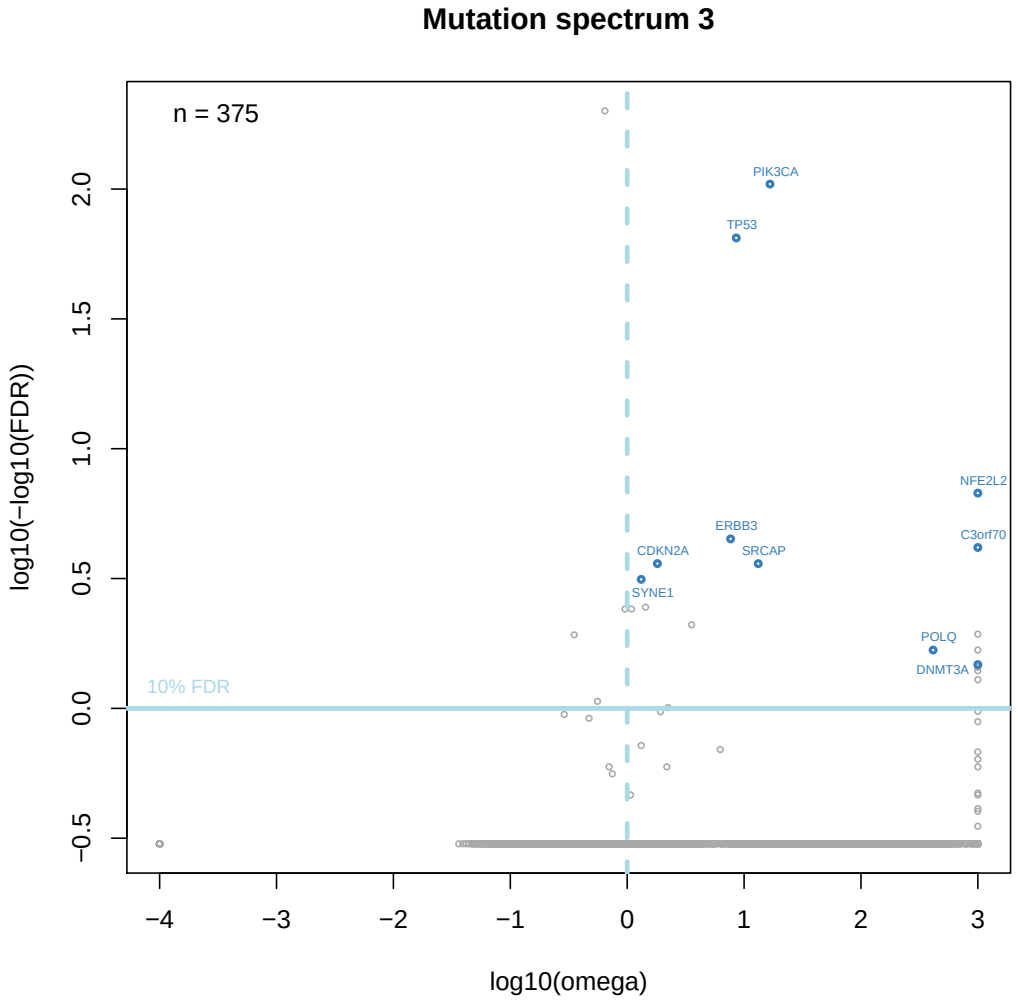


FIGURE 6.6: **Gene-based omega analysis in PAML for Signature 3.** For each gene in the group of 375 patients in Signature 3, the omega estimate and FDR value from PAML analysis has been plotted to show the patterns of selection across the dataset. *R code used to generate plot in Supplementary Appendix H.*

6.2.1.3 Signature 3

Signature 3 (Figure 6.2) contains 375 patients, over 17 of the 21 different cancer types in the Lawrence data. These patients have mostly C→T and C→G changes.

In total, 19 genes were found to be significantly mutated in Signature 3, tabulated in Table 6.6 with their omega estimate, p-values and gene descriptions.

TABLE 6.6: **Ranked list of significantly mutated genes in Signature 3.** Significantly mutated genes (omega>1 and FDR<0.1) from PAML analysis of Signature 3 have been tabulated in ascending order by p-value (descending order of significance). *Code used to generate table in Supplementary Appendix I.*

Gene	Omega	P-value	Description
PIK3CA	16.66	5.45e-109	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha
TP53	8.57	3.89e-69	tumor protein p53
NFE2L2	999.00	6.28e-11	nuclear factor, erythroid 2-like 2
ERBB3	7.67	1.40e-08	v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 3
C3orf70	999.00	3.59e-08	chromosome 3 open reading frame 70
CDKN2A	1.81	1.50e-07	cyclin-dependent kinase inhibitor 2A
SRCAP	13.22	1.72e-07	Snf2-related CREBBP activator protein
SYNE1	1.32	5.73e-07	spectrin repeat containing, nuclear envelope 1
NUP93	1.43	3.09e-06	nucleoporin 93kDa
FGFR3	1.09	4.07e-06	fibroblast growth factor receptor 3
KMT2C	3.55	9.10e-06	lysine (K)-specific methyltransferase 2C
FBXL13	999.00	1.44e-05	F-box and leucine-rich repeat protein 13
POLQ	413.82	3.01e-05	polymerase (DNA directed), theta
KLF5	999.00	3.14e-05	Kruppel-like factor 5 (intestinal)
DNMT3A	999.00	5.30e-05	DNA (cytosine-5-)-methyltransferase 3 alpha
SLC22A10	999.00	6.04e-05	solute carrier family 22, member 10
MAP1A	999.00	7.04e-05	microtubule-associated protein 1A
CASP8	999.00	9.47e-05	caspase 8, apoptosis-related cysteine peptidase
SETDB1	2.24	1.99e-04	SET domain, bifurcated 1

GO term analysis in GOrilla has produced the top ten processes enriched for the significant genes in this analysis (Table 6.7).

DNA polymerase theta (POLQ) has been detected as a significantly mutated gene in this particular mutational signature, with a very high omega value (413.82) suggesting that this gene is under positive selection in this subset of cancers, and a significant p-value (3.01e-05) supporting the model of positive selection in PAML.

POLQ is a gene that encodes the protein POLθ, an A-family DNA polymerase, and was first identified by Sharief et al. [1999]. DNA polymerases are enzymes that synthesise DNA in genome replication, but also play an important role in protecting the cell against DNA damage [Lange et al., 2011]. POLQ maps to chromosome 3 and contains

TABLE 6.7: **Enriched GO terms in Signature 3.** Top 10 enriched GO terms in Signature 3, using process ontology from GOrilla.

Process description	P-value	FDR q-value	Enrichment (N, B, n, b)
positive regulation of cell aging	1.23E-5	1.62E-1	357.02 (18208,6,17,2)
chromatin organisation	1.61E-5	1.06E-1	10.07 (18208,638,17,6)
regulation of neuron apoptotic process	2.36E-5	1.04E-1	22.79 (18208,188,17,4)
regulation of transcription from RNA polymerase II promoter	4.16E-5	1.37E-1	5.35 (18208,1602,17,8)
regulation of leukocyte apoptotic process	4.38E-5	1.15E-1	42.84 (18208,75,17,3)
replicative senescence	4.49E-5	9.85E-2	194.74 (18208,11,17,2)
regulation of neuron death	5.1E-5	9.59E-2	18.71 (18208,229,17,4)
positive regulation of muscle cell apoptotic process	5.39E-5	8.86E-2	178.51 (18208,12,17,2)
positive regulation of cellular amino acid metabolic process	6.22E-5	9.1E-2	17.78 (18208,241,17,4)
organelle organisation	6.28E-5	8.26E-2	4.29 (18208,2248,17,9)

an N-terminal ATP-binding domain and C-terminal DNA polymerase motifs A, B and C [Sharief et al., 1999].

This novel candidate cancer gene is of great interest, as it has not previously been associated with cancer. In the work of Heitzer and Tomlinson [2014], two other replicative DNA polymerases (POLE and POLD1) were originally discovered to be enriched with germline exonuclease domain mutations (EDMs) in human cancers predisposing to polymerase proofreading associated polyposis (PPAP), which is a disease characterised by multiple colorectal adenomas and carcinoma. POLE was also found to harbour somatic EDMs in sporadic colorectal and endometrial cancers. In tumours with EDMs, the 3' exonuclease proofreading activity is affected, causing an 'ultramutator' phenotype with an increase in base substitutions, but microsatellite stability is maintained. However POLQ was not investigated in this work.

Two papers recently published in Nature [Ceccaldi et al., 2015, Mateos-Gomez et al., 2015] investigated the role of POLQ in cancer and genome stability, presenting POLQ

as a druggable target in cancers with defective homologous recombination (HR) pathways. In Mateos-Gomez et al. [2015], next-generation sequencing showed that repair by alternative non-homologous end-joining (alt-NHEJ), also known as microhomology-mediated end-joining (MMEJ), yields non-TTAGGG nucleotide insertions at fusion breakpoints of dysfunctional telomeres. The enzymatic activity of POLQ was identified as being responsible for these random insertions, establishing POLQ as a crucial alt-NHEJ factor in mammalian cells. Additionally, in mice Mateos-Gomez et al. [2015] found that the inhibition of POLQ resulted in the suppression of alt-NHEJ at dysfunctional telomeres and hindered chromosomal translocations at non-telomeric loci, and loss of POLQ caused increased rates of homology-directed repair evident by recombination of dysfunctional telomeres and the accumulation of RAD51 at double-strand breaks. It was also shown that the depletion of POLQ had a synergistic effect on cell survival in the absence of BRCA1 or BRCA2, suggesting that, in tumours carrying mutations in homology-directed repair genes (such as BRCA1 or BRCA2), the inhibition of mutagenic POLQ could represent a valid therapeutic target. Supporting this notion, Ceccaldi et al. [2015] reported an inverse correlation between HR activity and POLQ expression in epithelial ovarian cancers (EOCs), with the knockdown of POLQ in HR-proficient cells upregulating HR activity and RAD51 nucleofilament assembly, while knockdown of POLQ in HR-deficient EOCs enhanced cell death. Furthermore, POLQ contains RAD51 binding motifs which block RAD51-mediated recombination, further supporting the role of POLQ as an inhibitor of HR. Consistent with the results in Mateos-Gomez et al. [2015], genetic inactivation of POLQ and the HR gene FANCD2 in mice resulted in embryonic lethality. Ceccaldi et al. [2015] concluded that the results revealed a synthetic lethal relationship between the HR pathway and POLQ-mediated repair in EOCs, with HR-deficient cells not able to survive in the absence of POLQ, again identifying POLQ as a novel druggable target in HR-defective tumours.

In both studies increased expression of POLQ was indicated to be driven by HR deficiency, in which the compensatory alt-NHEJ mechanism is promoted, and hence POLQ has been suggested as a synthetic lethal target in HR-deficient tumours. However driver mutations have not been identified in this gene in cancer, presenting POLQ as a novel

finding in this evolutionary analysis. Consistent with the findings from Mateos-Gomez et al. [2015] and Ceccaldi et al. [2015] that POLQ is overespressed in EOCs and other HR-deficient tumours [Ceccaldi et al., 2015] and upregulated in a wide range of human cancers [Mateos-Gomez et al., 2015], it is speculated that the putative driver mutations in POLQ in Signature 3 patients are activating mutations modifying the function of POLQ by upregulating POLQ expression so that alt-NHEJ is promoted and HR is suppressed, rather than inactivating mutations removing function. Since alt-NHEJ is more error prone than HR, and has been shown to lead to chromosomal translocations, it is predicted that activating mutations in POLQ contribute to genomic instability within tumours by inhibiting HR. Interestingly this gene has not been found to be significantly mutated in any cancer type in either the PAML analysis or the MutSig Lawrence analysis, suggesting that mutations in this gene are not tissue-specific but rather mutation spectrum dependent. The mutational profile observed in Signature 3 exhibits mostly C→A, C→G and C→T mutations, a signature typical of tumours with abnormalities in DNA maintenance, specifically defective HR-based DNA double-strand break repair [Alexandrov et al., 2013]. Signature 3 is also enriched for breast cancer, a tumour type known to be associated with mutations in genes involved in the maintenance of genome stability, such as BRCA1 and BRCA2 [Boulton, 2006]. Hence, Signature 3 could represent a mutational signature specific to HR-deficient tumours, supporting the role of POLQ as an oncogene associated with a mutation spectrum characterised by genomic instability, further validating POLQ as a potential therapeutic target in HR-deficient cancers.

POLQ was also suggested to maintain genomic stability at stalled or collapsed replication forks by promoting fork restart, in response to replication stress induced by UV light for example [Ceccaldi et al., 2015]. The mechanism for this is unknown, but was suggested to be mediated through interaction with HR, since RAD51 binding by POLQ was enhanced in cells under replicative stress. Therefore in tumours such as melanoma that are known to be induced by environmental exposures causing replicative stress, inactivating mutations in POLQ could potentially lead to the development of cancer

through increased abnormalities in replication fork progression. In this case POLQ could be acting as a tumour suppressor gene.

To further investigate the POLQ mutations that were detected in the subset of 375 patients in Signature 3, UniProt [Consortium et al., 2014] and InterPro [Mitchell et al., 2014] were used to identify domains and post-translational modification sites in POLQ, and Plot Protein [Turner, 2013] was used to visualise the locations of the mutations relative to functional regions. Of all 19 SNVs in POLQ over Signature 3 patients, only exonic mutations were used for this analysis, of which there were 15 unique mutations (all missense) over 13 different patients (Figure 6.7). The discarded four mutations were intronic variants, with three present in a single breast cancer (BRCA) and one in a lung squamous cell carcinoma (LUSC). The patients harbouring missense mutations were as follows: two LUSC; three endometrial (UCEC), with one patient containing two mutations; one BRCA; three bladder (BLCA), with one patient harbouring two mutations; and four head and neck cancers (HNSC). Both BRCA and UCEC, known HR-deficient associated tumour types, are amongst the cancer types affected by putative driver mutations in this particular signature, again supporting the role of POLQ as a cancer gene in HR-defective tumours.

Table 6.8 lists the amino acid changes highlighted in Figure 6.7, and shows which mutations are present in domains and post-translational modification sites.

Both Figure 6.7 and Table 6.8 show that seven of the POLQ missense SNVs called in the Signature 3 patients are located in POLQ domains, covering the helicase C-terminal, the DNA polymerase domain, the helicase ATP-binding domain (containing the DEAD box motif) and the ribonuclease H-like domain. In Ceccaldi et al. [2015], POLQ mutants lacking either ATPase catalytic activity or interaction with RAD51 were found to fully reduce RAD51 loci (increasing RAD51-ssDNA nucleofilament assembly) and increase the recombination frequency compared to wild-type POLQ, whereas a POLQ mutant lacking the polymerase domain was found to have the same effect on HR activity and RAD51 foci as the wild-type POLQ. This suggested that the N-terminal half of POLQ, containing the RAD51 binding domain and the ATPase domain, were sufficient for the

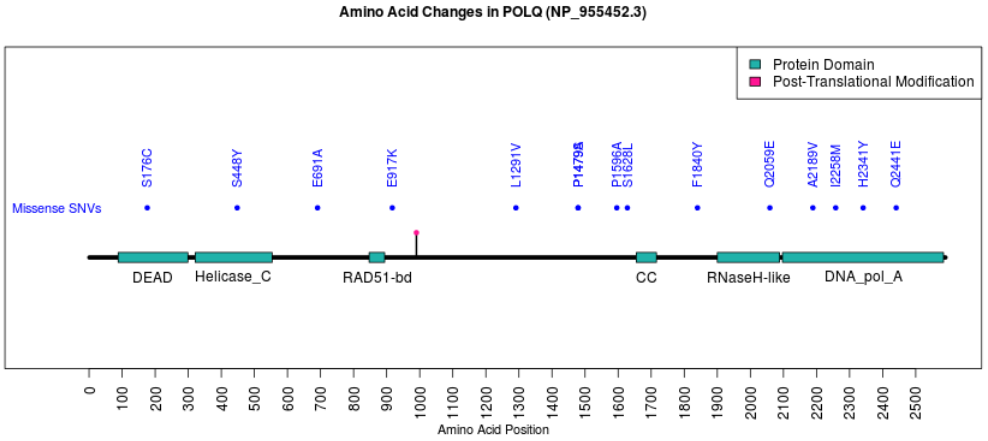


FIGURE 6.7: **Location of missense SNVs in POLQ within Signature 3.** Plot Protein has been used to visualise the locations of the 15 missense mutations called in POLQ within the Signature 3 subset containing 375 patients, relative to functional regions such as domains and post-translational modification sites. The plotted blue points denote the amino acid changes present in the gene, with the amino acid positions shown along the x-axis. The domains are highlighted in green with the following abbreviations: DEAD = helicase superfamily 1/2, ATP-binding domain (DEAD/ DEAH box helicase motif); Helicase_C = Helicase, C-terminal; RAD51-bd = RAD51 interacting site; CC = coiled coil; and DNA_pol_A = DNA-directed DNA polymerase, family A, palm domain. The post-translational modification site is highlighted in pink, and represents N6-acetyllysine at amino acid residue 990. POLQ protein length = 2590 amino acids.

disruption of RAD51 foci (by negatively regulating RAD51-ssDNA assembly) as well as the inhibition of HR. Conversely, the polymerase domain was found to be unnecessary for the regulation of these processes. Based on this, the mutation in amino acid 176 resulting in a S to C change in the helicase ATP-binding domain (DEAD box) found in the results of this analysis (Table 6.8) could potentially affect the function of POLQ by upregulating the suppression of HR, as was shown in Ceccaldi et al. [2015], resulting in the promotion of the more error-prone alt-NHEJ pathway which encourages a high genomic instability pattern (through an increased rate of chromosomal translocations) that is commonly observed in HR-deficient tumours. The polymerase domain of POLQ, in which point mutations were also detected in the PAML analysis, was shown to be required for the error-prone NHEJ pathway [Ceccaldi et al., 2015]. Therefore these particular mutations could be responsible for upregulating POLQ activity and promoting

TABLE 6.8: **Amino acid changes in POLQ within Signature 3.** The protein positions of all 15 amino acid changes in POLQ (protein ID NP_955452.3) within the Signature 3 subset containing 375 patients. The domains are abbreviated as follows: DNA-dir_DNA_pol_A_palm_dom = DNA-directed DNA polymerase family A, palm domain; Helicase_ATP-bd = Helicase superfamily 1/2, ATP-binding domain; and DEAD = DEAD/DEAH box helicase motif. AA = amino acid; PTM = post-translational modification.

AA position	Ref AA	Alt AA	Domain	PTM site
1479	P	S	Not in domain	No
1628	S	L	Not in domain	No
691	E	A	Not in domain	No
1596	P	A	Not in domain	No
1479	P	A	Not in domain	No
448	S	Y	Helicase C-terminal	No
2341	H	Y	DNA-dir_DNA_pol_A_palm_dom	No
2258	I	M	DNA-dir_DNA_pol_A_palm_dom	No
1840	F	Y	Not in domain	No
176	S	C	Helicase_ATP-bd (DEAD)	No
1291	L	V	Not in domain	No
12441	Q	E	DNA-dir_DNA_pol_A_palm_dom	No
2189	A	V	DNA-dir_DNA_pol_A_palm_dom	No
2059	Q	E	Ribonuclease H-like domain	No
917	E	K	Not in domain	No

alt-NHEJ to also lead to increased genomic instability and the subsequent development of cancer. No mutations were observed in the RAD51 binding domain in this analysis. However it has been suggested that the anti-recombinase activity of this domain is responsible for maintaining genomic stability in HR-defective tumours, presenting this region of POLQ as a good knockout target in tumours that are HR-deficient and also treated with PARPi (inhibition of PARP), PARP1 being another crucial factor in alt-NHEJ [Ceccaldi et al., 2015].

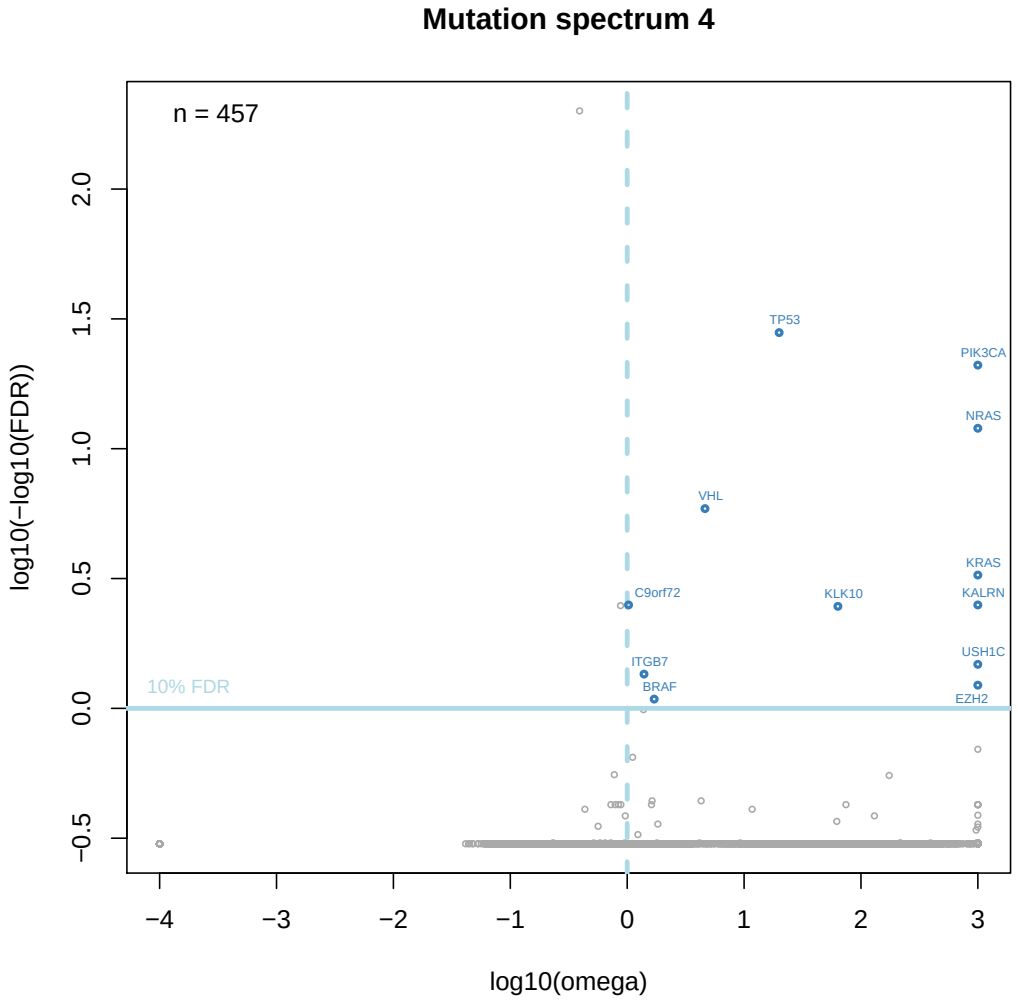


FIGURE 6.8: **Gene-based omega analysis in PAML for Signature 4.** For each gene in the group of 457 patients in Signature 4, the omega estimate and FDR value from PAML analysis has been plotted to show the patterns of selection across the dataset. *R code used to generate plot in Supplementary Appendix H.*

6.2.1.4 Signature 4

Signature 4 (Figure 6.2) contains 457 patients. These patients have a more uniform mutation spectra than the other signatures with fairly even proportions of each of the six single nucleotide changes.

TABLE 6.9: **Ranked list of significantly mutated genes in Signature 4.** Significantly mutated genes (omega>1 and FDR<0.1) from PAML analysis of Signature 4 have been tabulated in ascending order by p-value (descending order of significance). *Code used to generate table in Supplementary Appendix I.*

Gene	Omega	P-value	Description
TP53	19.94	2.32e-32	tumor protein p53
PIK3CA	999.00	3.58e-25	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha
NRAS	999.00	4.76e-16	neuroblastoma RAS viral (v-ras) oncogene homolog
VHL	4.63	7.65e-10	von Hippel-Lindau tumor suppressor, E3 ubiquitin protein ligase
KRAS	999.00	3.81e-07	Kirsten rat sarcoma viral oncogene homolog
KALRN	999.00	2.73e-06	kalirin, RhoGEF kinase
C9orf72	1.03	2.92e-06	chromosome 9 open reading frame 72
KLK10	63.49	3.91e-06	kallikrein-related peptidase 10
USH1C	999.00	4.25e-05	Usher syndrome 1C (autosomal recessive, severe)
ITGB7	1.39	6.15e-05	integrin, beta 7
EZH2	999.00	8.91e-05	enhancer of zeste homolog 2 (Drosophila)
BRAF	1.70	1.33e-04	v-raf murine sarcoma viral oncogene homolog B

In total, 12 genes were found to be significantly mutated in Signature 4, tabulated in Table 6.9 with their omega estimate, p-values and gene descriptions.

GO term analysis in GOrilla has produced the top ten processes enriched for the significant genes in this analysis (Table 6.10).

TABLE 6.10: **Enriched GO terms in Signature 4.** Top 10 enriched GO terms in Signature 4, using process ontology from GOrilla.

Process description	P-value	FDR q-value	Enrichment (N, B, n, b)
regulation of neuron apoptotic process	8.31E-8	1.09E-3	40.35 (18208,188,12,5)
regulation of neuron death	2.22E-7	1.46E-3	33.13 (18208,229,12,5)
neurotrophin TRK receptor signaling pathway	5.6E-7	2.46E-3	27.49 (18208,276,12,5)
neurotrophin signaling pathway	6.02E-7	1.98E-3	27.10 (18208,280,12,5)
negative regulation of neuron apoptotic process	1.21E-6	3.19E-3	46.33 (18208,131,12,4)
negative regulation of neuron death	2.37E-6	5.2E-3	39.16 (18208,155,12,4)
positive regulation of protein phosphorylation	2.41E-6	4.54E-3	12.99 (18208,701,12,6)
fibroblast growth factor receptor signaling pathway	3.12E-6	5.13E-3	36.56 (18208,166,12,4)
visual learning	3.72E-6	5.44E-3	94.83 (18208,48,12,3)
positive regulation of Rac protein signal transduction	3.98E-6	5.24E-3	606.93 (18208,5,12,2)

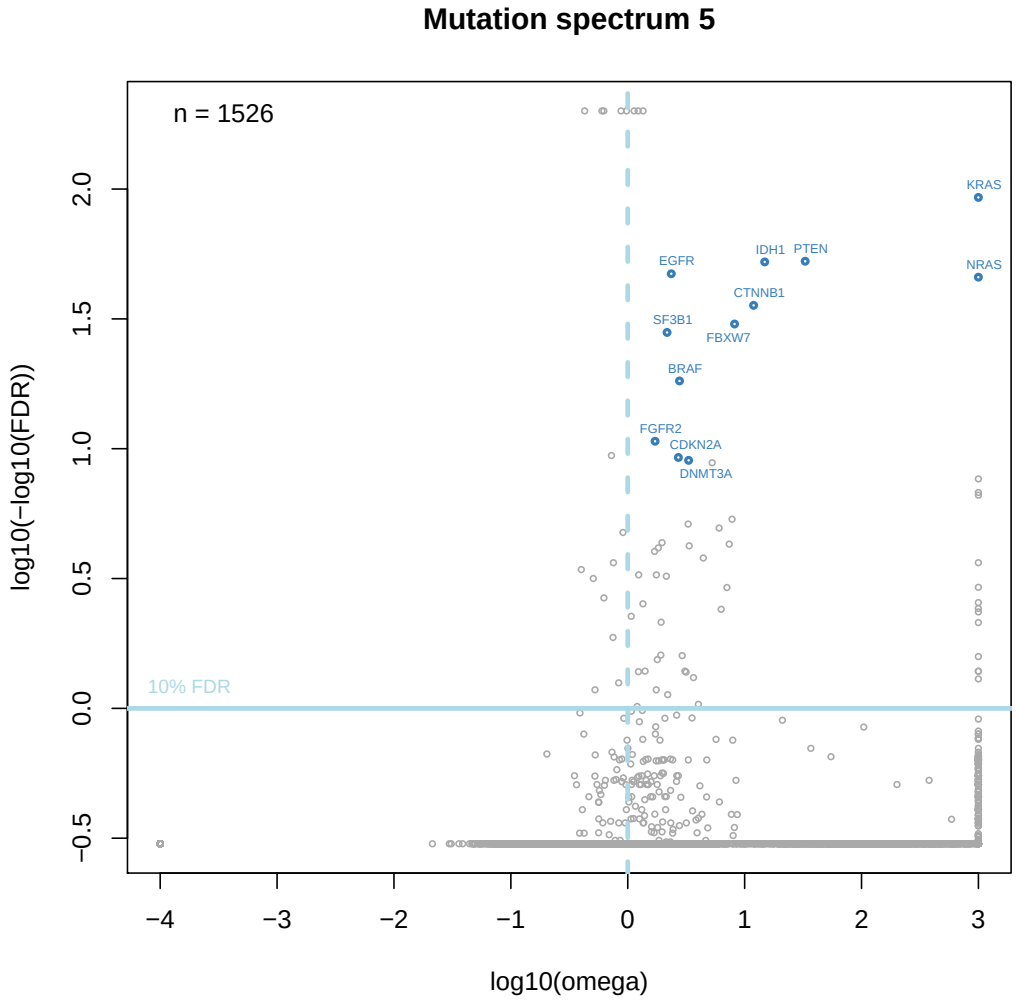


FIGURE 6.9: **Gene-based omega analysis in PAML for Signature 5.** For each gene in the group of 1,526 patients in Signature 5, the omega estimate and FDR value from PAML analysis has been plotted to show the patterns of selection across the dataset. *R code used to generate plot in Supplementary Appendix H.*

6.2.1.5 Signature 5

Signature 5 (Figure 6.2) contains 1,526 patients. These patients have mostly C→T changes, but this signature differs from Signature 2 as it has a much lower rate of these changes reaching ∼60% whereas in Signature 2 the C→T rate reaches almost 100% in some patients.

In total, 58 genes were found to be significantly mutated in Signature 5, and the top most significant 35 of these have been tabulated in Table 6.11 with their omega estimate, p-values and gene descriptions.

GO term analysis in GOrilla has produced the top ten processes enriched for the significant genes in this analysis (Table 6.12).

TABLE 6.11: **Ranked list of significantly mutated genes in Signature 5.** Significantly mutated genes (omega>1 and FDR<0.1) from PAML analysis of Signature 5 have been tabulated in ascending order by p-value (descending order of significance). List has been truncated to n=35 rows out of a total of 58 significant genes. *Full table in Supplementary Appendix J. Code used to generate table in Supplementary Appendix I.*

Gene	Omega	P-value	Description
SYNE1	1.23	0.00e+00	spectrin repeat containing, nuclear envelope 1
FLG	1.35	0.00e+00	filaggrin
LRP1B	1.13	0.00e+00	low density lipoprotein receptor-related protein 1B
KRAS	999.00	8.10e-97	Kirsten rat sarcoma viral oncogene homolog
PTEN	32.95	1.20e-56	phosphatase and tensin homolog
IDH1	14.85	2.85e-56	isocitrate dehydrogenase 1 (NADP+), soluble
EGFR	2.36	5.14e-51	epidermal growth factor receptor
NRAS	999.00	1.41e-49	neuroblastoma RAS viral (v-ras) oncogene homolog
CTNNB1	11.93	2.02e-39	catenin (cadherin-associated protein), beta 1, 88kDa
FBXW7	8.20	5.95e-34	F-box and WD repeat domain containing 7, E3 ubiquitin protein ligase
SF3B1	2.17	9.47e-32	splicing factor 3b, subunit 1, 155kDa
BRAF	2.77	6.42e-22	v-raf murine sarcoma viral oncogene homolog B
FGFR2	1.71	2.45e-14	fibroblast growth factor receptor 2
CDKN2A	2.71	7.33e-13	cyclin-dependent kinase inhibitor 2A
DNMT3A	3.32	1.32e-12	DNA (cytosine-5-)-methyltransferase 3 alpha
MYD88	5.28	2.13e-12	myeloid differentiation primary response 88
PIK3R1	999.00	3.39e-11	phosphoinositide-3-kinase, regulatory subunit 1 (alpha)
IDH2	999.00	2.61e-10	isocitrate dehydrogenase 2 (NADP+), mitochondrial
SMAD4	999.00	3.91e-10	SMAD family member 4
VHL	7.81	7.62e-09	von Hippel-Lindau tumor suppressor, E3 ubiquitin protein ligase
U2AF1	3.29	1.33e-08	U2 small nuclear RNA auxiliary factor 1
CTCF	6.05	2.07e-08	CCCTC-binding factor (zinc finger protein)
RUNX1	1.96	8.86e-08	runt-related transcription factor 1
SPOP	7.38	1.04e-07	speckle-type POZ protein
DCHS1	3.35	1.25e-07	dachsous cadherin-related 1
CDC27	1.83	1.54e-07	cell division cycle 27
AKT1	1.70	2.12e-07	v-akt murine thymoma viral oncogene homolog 1
CELSR3	4.43	3.70e-07	cadherin, EGF LAG seven-pass G-type receptor 3
ZDHHC4	999.00	5.59e-07	zinc finger, DHHC-type containing 4
DNAH2	1.24	1.39e-06	dynein, axonemal, heavy chain 2
PPP2R1A	1.76	1.42e-06	protein phosphatase 2, regulatory subunit A, alpha
KMT2D	2.14	1.59e-06	lysine (K)-specific methyltransferase 2D
TMEM109	999.00	3.36e-06	transmembrane protein 109
DCDC1	7.05	3.48e-06	doublecortin domain containing 1
TCEB1	999.00	8.43e-06	transcription elongation factor B (SIII), polypeptide 1 (15kDa, elongin C)

TABLE 6.12: **Enriched GO terms in Signature 5.** Top 10 enriched GO terms in Signature 5, using process ontology from GOrilla.

Process description	P-value	FDR q-value	Enrichment (N, B, n, b)
fibroblast growth factor receptor signaling pathway	1.08E-10	1.42E-6	18.91 (18208,166,58,10)
positive regulation of macromolecule metabolic process	2.76E-9	1.82E-5	3.44 (18208,2372,58,26)
positive regulation of gene expression	6.95E-9	3.05E-5	4.40 (18208,1428,58,20)
positive regulation of cellular metabolic process	8.73E-9	2.87E-5	3.26 (18208,2503,58,26)
neurotrophin TRK receptor signaling pathway	1.5E-8	3.94E-5	11.37 (18208,276,58,10)
positive regulation of nitrogen compound metabolic process	1.56E-8	3.43E-5	3.97 (18208,1661,58,21)
neurotrophin signaling pathway	1.72E-8	3.23E-5	11.21 (18208,280,58,10)
positive regulation of signaling	1.87E-8	3.07E-5	4.40 (18208,1356,58,19)
positive regulation of cell communication	2.05E-8	3E-5	4.37 (18208,1364,58,19)
positive regulation of signal transduction	2.72E-8	3.58E-5	4.58 (18208,1235,58,18)

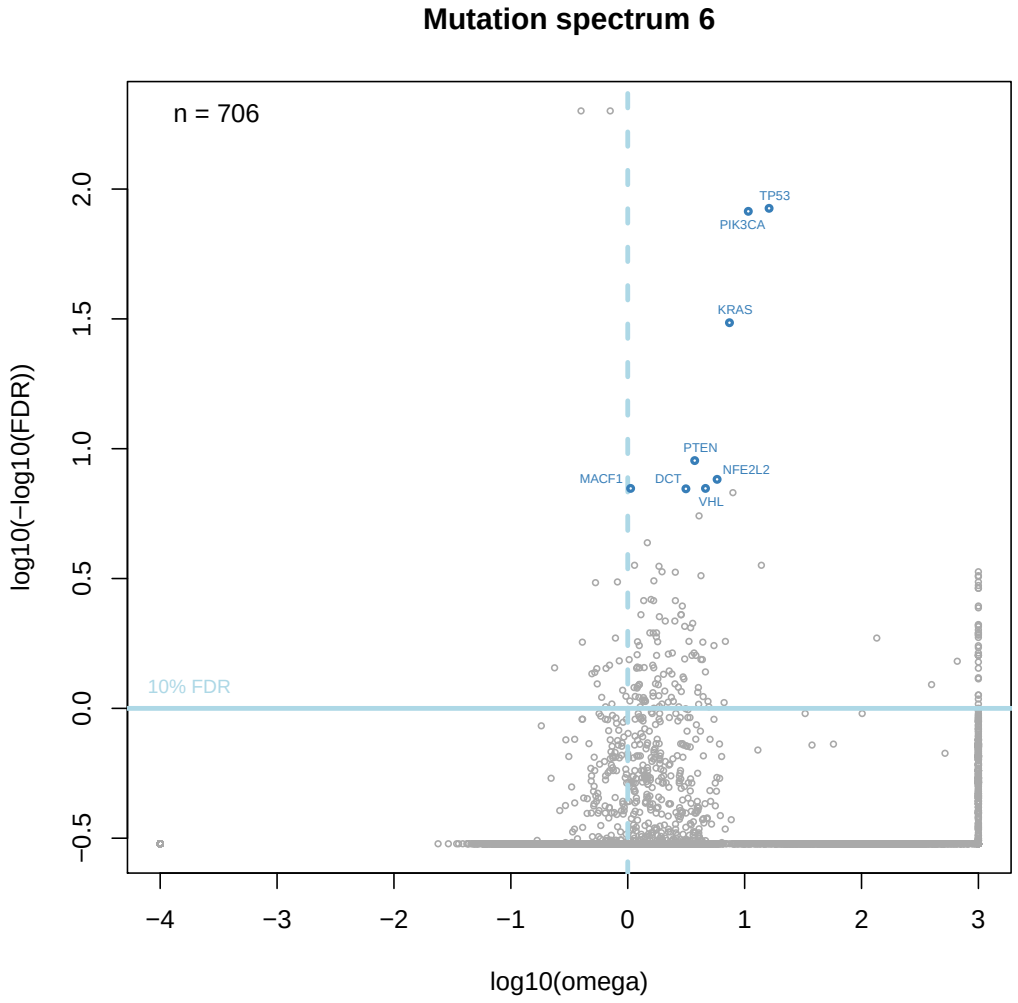


FIGURE 6.10: **Gene-based omega analysis in PAML for Signature 6.** For each gene in the group of 706 patients in Signature 6, the omega estimate and FDR value from PAML analysis has been plotted to show the patterns of selection across the dataset. *R code used to generate plot in Supplementary Appendix H.*

6.2.1.6 Signature 6

Signature 6 (Figure 6.2) contains 706 patients. These patients have mostly C→T changes again, but within the other five types of change C→A and C→G are slightly elevated above T→A, T→C and T→G which are all more uniform.

In total, 141 genes were found to be significantly mutated in Signature 6, and the top most significant 35 of these have been tabulated in Table 6.13 with their omega estimate, p-values and gene descriptions.

GO term analysis in GOrilla has produced the top ten processes enriched for the significant genes in this analysis (Table 6.14).

TABLE 6.13: **Ranked list of significantly mutated genes in Signature 6.** Significantly mutated genes (omega>1 and FDR<0.1) from PAML analysis of Signature 6 have been tabulated in ascending order by p-value (descending order of significance). List has been truncated to n=35 rows out of a total of 141 significant genes. *Full table in Supplementary Appendix J. Code used to generate table in Supplementary Appendix I.*

Gene	Omega	P-value	Description
TP53	16.18	9.72e-89	tumor protein p53
PIK3CA	10.75	2.20e-86	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha
KRAS	7.42	8.81e-35	Kirsten rat sarcoma viral oncogene homolog
PTEN	3.74	3.93e-13	phosphatase and tensin homolog
NFE2L2	5.81	1.10e-11	nuclear factor, erythroid 2-like 2
MACF1	1.06	5.01e-11	microtubule-actin crosslinking factor 1
VHL	4.63	5.49e-11	von Hippel-Lindau tumor suppressor, E3 ubiquitin protein ligase
DCT	3.14	6.45e-11	dopachrome tautomerase
PIK3R1	7.93	1.22e-10	phosphoinositide-3-kinase, regulatory subunit 1 (alpha)
APC	4.07	2.43e-09	adenomatous polyposis coli
KIAA2026	1.47	3.87e-08	KIAA2026
DMXL2	13.91	2.72e-07	Dmx-like 2
LRP1B	1.14	2.74e-07	low density lipoprotein receptor-related protein 1B
ZNF749	1.85	3.14e-07	zinc finger protein 749
UVRAG	999.00	5.08e-07	UV radiation resistance associated
XPO1	1.96	5.20e-07	exportin 1 (CRM1 homolog, yeast)
WWTR1	2.55	5.66e-07	WW domain containing transcription regulator 1
ZNF283	999.00	7.77e-07	zinc finger protein 283
ARID1A	4.23	7.90e-07	AT rich interactive domain 1A (SWI-like)
ZNHIT6	999.00	8.82e-07	zinc finger, HIT-type containing 6
KMT2D	1.67	1.21e-06	lysine (K)-specific methyltransferase 2D
ASB15	999.00	1.37e-06	ankyrin repeat and SOCS box containing 15
CEP89	999.00	1.90e-06	centrosomal protein 89kDa
RBL2	999.00	2.37e-06	retinoblastoma-like 2 (p130)
STIM2	999.00	2.40e-06	stromal interaction molecule 2
CAPRIN2	1.58	4.67e-06	caprin family member 2
KCND3	1.37	5.13e-06	potassium voltage-gated channel, Shal-related subfamily, member 3
DYM	2.56	5.34e-06	dymeclin
TOX	1.66	5.47e-06	thymocyte selection-associated high mobility group box
NRAS	2.93	7.45e-06	neuroblastoma RAS viral (v-ras) oncogene homolog
PDZD8	999.00	7.66e-06	PDZ domain containing 8
APOLD1	999.00	8.63e-06	apolipoprotein L domain containing 1
FBXW7	2.87	1.29e-05	F-box and WD repeat domain containing 7, E3 ubiquitin protein ligase
AP000295.9	2.85	1.30e-05	
NEB	1.30	1.30e-05	nebulin

TABLE 6.14: **Enriched GO terms in Signature 6.** Top 10 enriched GO terms in Signature 6, using process ontology from GOrilla.

Process description	P-value	FDR q-value	Enrichment (N, B, n, b)
fibroblast growth factor receptor signaling pathway	4.63E-6	6.09E-2	7.21 (18208,166,137,9)
regulation of phosphorylation	3.79E-5	2.49E-1	2.51 (18208,1218,137,23)
neurotrophin TRK receptor signaling pathway	4.6E-5	2.02E-1	4.82 (18208,276,137,10)
neurotrophin signaling pathway	5.19E-5	1.71E-1	4.75 (18208,280,137,10)
enzyme linked receptor protein signaling pathway	6.25E-5	1.64E-1	2.84 (18208,843,137,18)
regulation of protein targeting	6.34E-5	1.39E-1	5.94 (18208,179,137,8)
positive regulation of response to stimulus	6.65E-5	1.25E-1	2.17 (18208,1715,137,28)
regulation of phosphorus metabolic process	6.78E-5	1.12E-1	2.30 (18208,1442,137,25)
regulation of intracellular protein transport	7.06E-5	1.03E-1	5.11 (18208,234,137,9)
Fc-epsilon receptor signaling pathway	7.7E-5	1.01E-1	5.78 (18208,184,137,8)

6.2.2 Measuring and estimating run times in PAML

Time constraints meant that the data running through PAML for this mutation spectra sub-type analysis had to be reorganised to maximise and prioritise the efficiency of the analysis.

Using the Lawrence dataset over all 4,728 patients and 21 cancer types, the run time was measured for each gene that had successfully completed ten iterations in PAML, in order to determine how long the genes in the mutation spectra sub-analysis would take to complete analysis.

To ascertain whether number of patient alignments per gene or length of gene was the more time-limiting factor in this analysis, the relationship between time taken and each of these variables was measured and plotted in Figure 6.11 and Figure 6.12.

Figure 6.11 and Figure 6.12 show that run time in PAML is more correlated with patient number rather than gene length, showing a strong positive correlation. Therefore, the number of patients is a more important consideration when estimating how long each gene will take to run through PAML.

Over the whole Lawrence dataset, TTN had the most number of patients (1400). However in this mutation spectra sub-type analysis, the gene with the largest number of patients was TP53 in Signature 5 with 406 patients.

In Figure 6.13, a linear regression (Equation 6.1) has been fitted to the data from Figure 6.11 and the straight line of best fit has been extrapolated to predict how long TP53 will take to run through PAML ten times. According to the equation of the straight line, TP53 with 406 patient alignments is estimated to take ∼40 hours.

$$y = 395.913x - 17255.573 \tag{6.1}$$

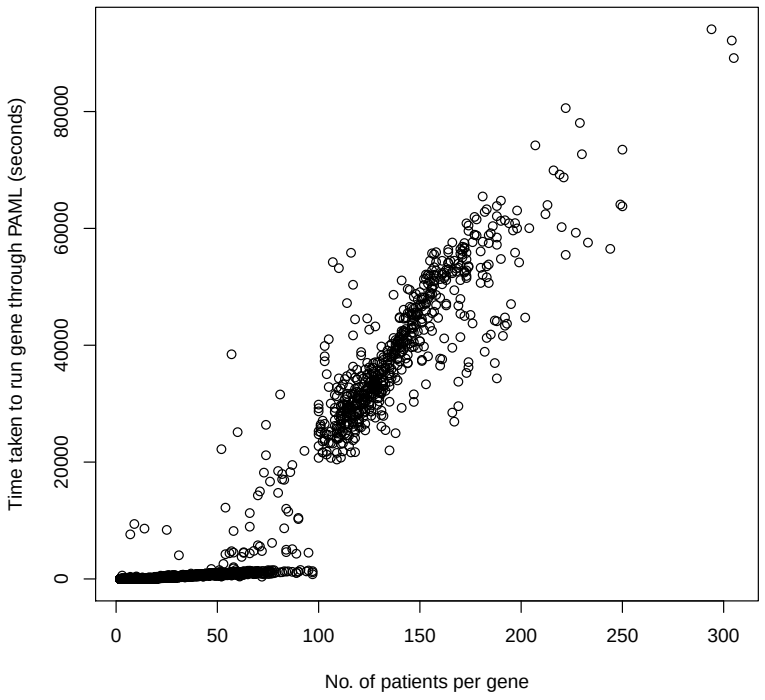


FIGURE 6.11: **Relationship between patient number and PAML run time.** For each gene in the whole Lawrence dataset of 4,728 patients that had successfully run though PAML to completion, the number of patients with mutations in that gene was plotted along the x-axis against the time it took (seconds) to run the gene through PAML along the y-axis, to show relationship between these two variables.

6.3 Discussion

The Lawrence dataset has been stratified by mutation spectra in order to detect signals of positive selection in candidate cancer driver genes specific to certain types of mutation spectra. Signature 3 has revealed a potential candidate cancer gene.

6.3.1 Relating mutational profile to path of selection

Many of the same significantly mutated genes are found across almost all six mutational signatures, such as TP53 and PIK3CA, however there are also genes such as POLQ

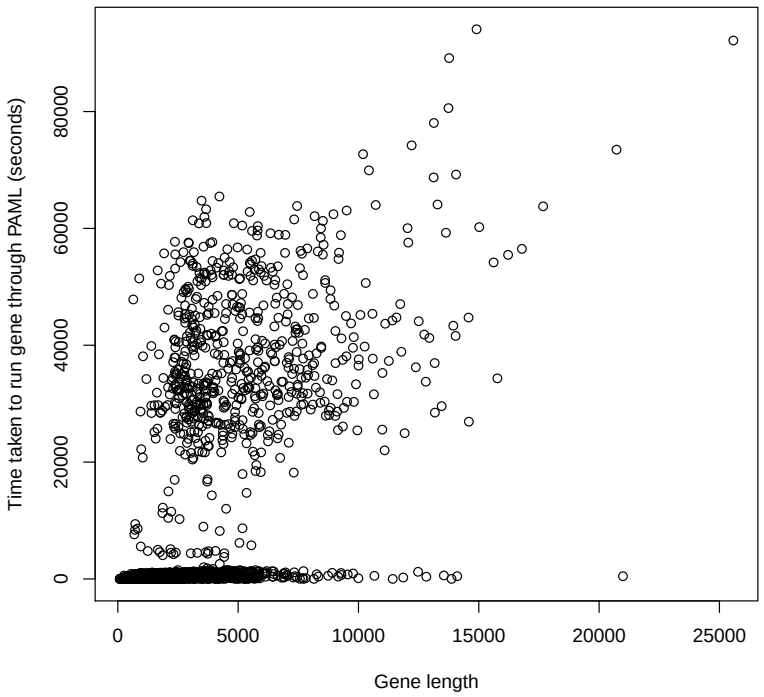


FIGURE 6.12: **Relationship between gene length and PAML run time.** For each gene in the whole Lawrence dataset of 4,728 patients that had successfully run through PAML, the gene length was plotted along the x-axis against the time it took (seconds) to run the gene through PAML along the y-axis, to show the relationship between these two variables.

that are unique to a specific mutation spectra. Therefore it is worth accounting for the mutation spectra of a cancer since it seems that the profile of a cancer may affect the genes that are hit by driver mutations in cancer.

However, the classification system used in this analysis is fairly basic, and should be further refined to include sequence context and CpG effects for example using the classification systems of Alexandrov et al. [2013] and Kandoth et al. [2013], in order to separate out the mutation spectra further. The way it has been done here, the mutation spectra look quite similar in some cases between different signatures which may be why the same significant genes are seen in several of the signatures.

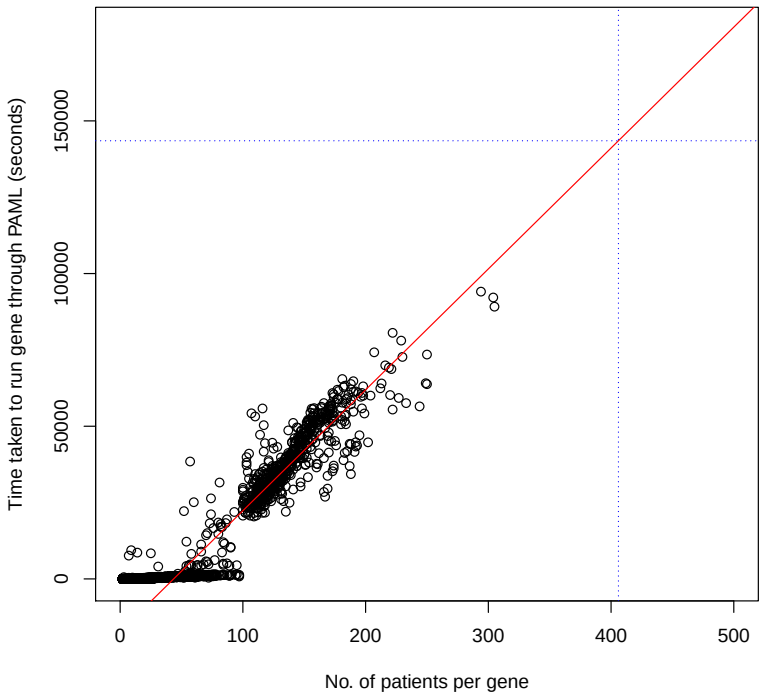


FIGURE 6.13: **Extrapolation for PAML run time estimates.** For each gene, the number of patients with mutations in that gene has been plotted along the x-axis against the time taken (seconds) to run the gene through PAML along the y-axis. A linear regression (red straight line of best fit) has been fitted to this data and extrapolated to estimate the PAML run time for the maximum patient number of 406 in TP53, which is estimated to take 143485 seconds (blue dotted lines).

6.3.2 Candidate cancer gene in Signature 3: POLQ

POLQ has been found to be significantly mutated in Signature 3. This is a very interesting finding, since it has not come up as significant in any of the tissue of origin sub-type analyses, and is only significant in Signature 3. This suggests that there is merit in grouping the data in this way, as otherwise this gene would not have reached significance.

POLQ has not yet been confirmed as a cancer gene, however it has been recently presented as a druggable target in cancers with defective homologous recombination [Ceccaldi et al., 2015, Mateos-Gomez et al., 2015], which makes this gene a good candidate cancer gene.

6.4 Methods

6.4.1 Partitioning data by single nucleotide mutation patterns

The Lawrence data was stratified by mutation spectra in the following way:

- Over all 4,728 patients in the Lawrence dataset, the relative proportions of each of the six classes of single nucleotide change (C→A, C→G, C→T, T→A, T→C, T→G) were calculated for each patient.
- A Euclidean (default in R) distance matrix was created in R containing all six relative single nucleotide change proportions for each patient.
- Cluster analysis was performed on this matrix using hclust in R, which scores the similarities between each patient, to group together patients with similar mutation spectra. This was done using hierarchical, agglomerative clustering using complete linkage (furthest-neighbour joining).
- The cluster analysis scores were used to create a cluster dendrogram in R to visualise the clustering, where the patients were clustered together based on their similarity scores (similarities in mutation spectra).
- From this dendrogram, an arbitrary height cut-off of branch length 0.65 was chosen to split the patients into 18 clusters, and six main groups were chosen from this.

6.4.2 PAML analysis

For each mutational signature, patient alignments edited with the Lawrence cancer-specific mutations were run through PAML in the same way as was done for the tissue of origin subsets, on a per gene basis as before.

An omega plot was generated in R for each of the six mutational signatures in the same way as has been done for the tissue of origin sub-type analysis in Chapter 5. Tables were also created listing the most significant genes for each mutational signature.

6.4.3 Gene ontology analysis

GOrilla (Gene ontology enrichment analysis and visualisation tool) was used to identify and visualise enriched GO terms in the six different mutational signatures [Eden et al., 2009].

GOrilla can be run in two modes, either using a single ranked list of genes as input or two unranked lists of genes (a target list and a background list). GOrilla was used here to search for enriched GO terms in a target list of genes compared to a background list of genes. For each mutational signature, the target list provided was the unranked list of significant genes (omega>1 and FDR<0.1) for that group and the background list was an unranked list of all 63677 genes obtained from Ensembl 78.

Gorilla is able to search for various ontologies: process, function or component. In this case process was the ontology searched for.

GOrilla calculated three values for each analysis:

- A **p-value** that is not corrected for multiple testing of 13167 GO terms. These were used to rank the process ontologies (smallest p-value most statistically significant).

- A **FDR q-value** which is the p-value corrected for multiple testing using the Benjamini and Hochberg method [Benjamini and Hochberg, 1995], so for the i^{th} term the FDR q-value (ranked by p-value) is (p-value * number of GO terms)/i.

- An **enrichment score** of (b/n)/(BN), *where N = the total number of genes (18208 genes from background list that were recognised, unique (highest ranking instance of each duplicated gene kept) and associated with a GO term), B = the total number of genes associated with a specific GO term, n = the number of genes in the target set and b = the number of genes in the intersection.*

Chapter 7

Mutation profile of FARP genes

7.1 Introduction

Following on from the finding in Chapter 3 that a subset of 16 glioblastoma multiforme (GBM) patients exhibited an increased rate of called INDELs, and building on the SNV analysis of varying mutation spectra in Chapter 3, the TCGA dataset was mined for specific mutation spectra characterised by high rates of called INDELs.

Both FARP1 and FARP2 genes were found to exhibit a strikingly high rate of cancer-specific INDELs in the TCGA dataset. These genes have also been identified as mechanistically novel regulators of autophagy in cancer cells (personal communication with Simon Wilkinson, IGMM Edinburgh). Neither FARP1 nor FARP2 have been causally implicated in cancer [Forbes et al., 2011]. Hence, it was speculated that these genes could be important functional candidate cancer genes, undergoing an interesting mutation spectrum defined by a high rate of INDELs.

FERM, RhoGEF, and pleckstrin homology domain protein 2 (FARP2) and its close homolog FERM, RhoGEF, and pleckstrin homology domain protein 1 (FARP1) are large multi-domain proteins sharing the same domain structures: FERM, RhoGEF,

and pleckstrin homology domains. FARP2 is also known as FERM domain including RhoGEF (FIR) and FGD1-related Cdc42-GEF (FRG), and FARP1 is also known as chondrocyte-derived ezrin-like protein (CDEP) and pleckstrin homology domain-containing family C member 2 (PLEKHC2). Functional studies of FARP1 and FARP2 have previously focused primarily on their roles in the regulation of neuronal development and morphology, motivated by their abundant expression both in neurons at the developmental stage and in the adult brain [He et al., 2013].

Interestingly, FARP1 and FARP2 are not shown to be significantly mutated in the Lawrence et al. [2014] study, nor in the PAML analysis carried out in this project. This suggests that these genes do not harbour driver SNVs, and so if they are causally implicated in cancer they are likely to be altered by another type of genetic mutation.

This chapter seeks to investigate the high rate of INDELs called in FARP1 and FARP2, to further understand their potential role in cancer.

7.2 Results

7.2.1 Initial COSMIC analysis of somatic mutations in FARP genes

Using the resource COSMIC [Forbes et al., 2010] to investigate acquired mutations in FARP1 and FARP2 in cancer, it was found that out of a total of 22,894 unique samples, 214 were mutated (0.77%) in FARP1. And out of a total number of 22,770 unique samples, 167 were mutated (0.73%) in FARP2. This included 12 mutant samples with substitution nonsense mutations, 131 with substitution missense, 68 with substitution synonymous, 1 with an insertion frameshift, 1 with deletion inframe and 3 with deletion frameshifts in FARP1. In FARP2, 9 mutant samples had substitution nonsense mutations, 96 had substitution missense, 49 had substitution synonymous, 3 had insertion frameshift, 1 had a deletion inframe and 5 had a deletion frameshift. These numbers do not necessarily equal the total number of mutated samples for each gene, as a single

sample can have mutations in one or more categories, but will only be counted once in the total number of mutated samples.

However, a limitation of using COSMIC to identify mutated genes in cancer is that it presents an acquisition bias, making it difficult to score the significance of genes. Mutations are recorded in COSMIC based on what has been found without accounting for the different types of genes that have preferentially been searched. For example more kinase genes will have been investigated due to their role in cancer, so there will be a bias towards more mutations represented in these genes.

7.2.2 FARP SNVs in TCGA dataset

The mySQL database containing all called variants in the TCGA dataset was mined to find SNVs targeting FARP1 and FARP2.

Over the whole TCGA dataset of 1,005 patients, FARP1 and FARP2 were shown to be hit by 20 and 30 cancer-specific coding heterozygous SNVs respectively, corresponding to a mean of 0.02 SNVs per patient in FARP1 and 0.03 in FARP2. Table 7.1 and Table 7.2 show how the mutations are spread over the different cancer types in both FARP1 and FARP2 respectively. In both genes, most mutations are found in the GBM dataset. However after accounting for varying patients numbers between cancer type subsets, of the seven cancer types that contain cancer-specific heterozygous SNV mutations in FARP1 it is brain lower grade glioma (LGG) that is the most enriched with SNVs (0.06), and of the eight cancer types hit by cancer-specific heterozygous SNVs in FARP2 GBM remains the cancer type exhibiting the highest mutation rate in this gene (0.09). Interestingly, the cancer types that are most enriched with FARP1 and FARP2 SNVs are both cancers of the brain, which is where these genes are known to be abundantly expressed [He et al., 2013].

Of the coding SNVs reported, most are non-synonymous. In FARP1, 80% are non-synonymous (Table 7.3), and in FARP2, 63% are non-synonymous (Table 7.4).

TABLE 7.1: **FARP1 cancer-specific SNV counts by tumour type.** For each tumour type, the cancer-specific heterozygous coding SNV counts over all 1005 patients in the TCGA dataset have been tabulated for FARP1, with the mean number of SNVs per patient for each dataset (calculated to 3dp).

Tumour type	SNV count	Mean number of SNVs per patient
BRCA	1	0.009
GBM	10	0.048
HNSC	1	0.012
KIRC	2	0.011
LGG	3	0.060
LUSC	2	0.038
UCEC	1	0.026
Whole dataset	**20**	**0.020**

TABLE 7.2: **FARP2 cancer-specific SNV counts by tumour type.** For each tumour type, the cancer-specific heterozygous coding SNV counts over all 1005 patients in the TCGA dataset have been tabulated for FARP2, with the mean number of SNVs per patient for each dataset (calculated to 3dp).

Tumour type	SNV count	Mean number of SNVs per patient
CESC	1	0.071
GBM	19	0.091
HNSC	2	0.024
KIRC	1	0.006
LGG	1	0.020
LUSC	1	0.019
OV	4	0.053
UCEC	1	0.026
Whole dataset	**30**	**0.030**

TABLE 7.3: **FARP1 cancer-specific SNV counts by variant consequence.** For each coding variant consequence, the cancer-specific heterozygous coding SNV counts over all 1005 patients in the TCGA dataset have been tabulated for FARP1.

Variant consequence	SNV count
NON-SYNONYMOUS	16
SYNONYMOUS	4
Total	**20**

TABLE 7.4: **FARP2 cancer-specific SNV counts by variant consequence.** For each coding variant consequence, the cancer-specific heterozygous coding SNV counts over all 1005 patients in the TCGA dataset have been tabulated for FARP2.

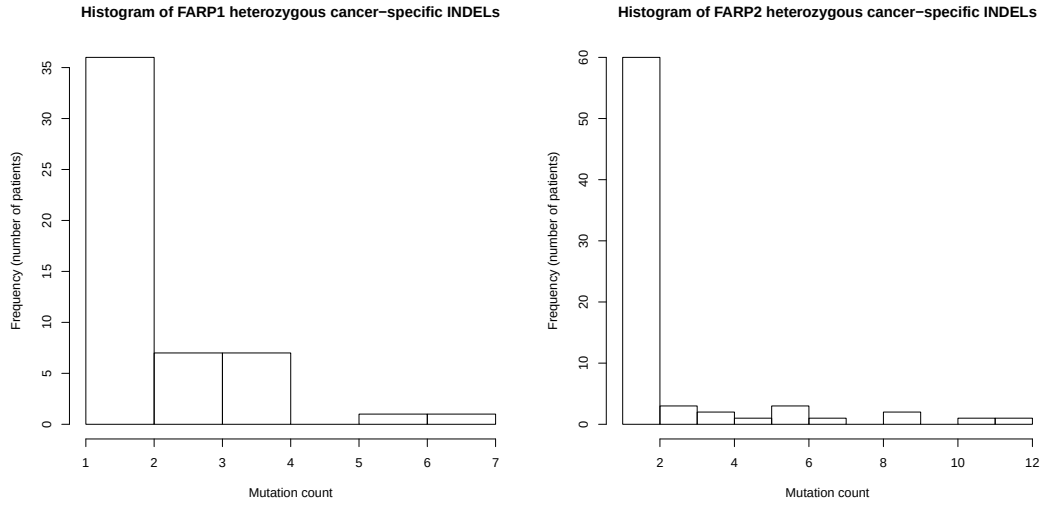
Variant consequence	SNV count
NON-SYNONYMOUS	19
SYNONYMOUS	11
Total	**30**

7.2.3 FARP INDELs in TCGA dataset

In the TCGA dataset of 1,005 patients, 52 patients were found to contain at least one heterozygous coding cancer-specific INDEL in FARP1, and 74 were found to contain at least one heterozygous coding cancer-specific INDEL in FARP2. The distribution of INDELs among these patients is shown in Figure 7.1, in which it can be seen that the distribution is more spread in FARP2 with more patients containing slightly higher rates of INDELs. However overall the shape of the distributions are largely the same with both patients in each gene exhibiting much smaller numbers of INDELs (i.e. one or two per patient). The patient with the highest rate of INDELs in FARP1 contains seven INDELs and is of the tumour type GBM. Of the 52 patients mutated with INDELs in FARP1, 36 are GBM (69%). The patient with the most INDELs in FARP2 contains 12 INDELs and has the tumour type OV. Of the 74 patients mutated with INDELs in FARP2, 17 are OV (23%).

It can be seen from Table 7.5, Table 7.6, Table 7.7 and Table 7.8 that INDEL rates are much higher than SNV rates in FARP1 and FARP2, with 109 INDELs compared to just 20 SNVs found in FARP1 across the whole TCGA dataset of 1,005 patients and 157 found in FARP2 compared to just 30 SNVs in this gene.

GBM generally has the highest mean rate of cancer-specific INDELs per patient in FARP1 with a mean number of 0.40 INDELs per patient, followed by OV with the second highest rate of 0.25 INDELs per patient (Table 7.5). The average number of FARP1 INDELs per patient for the TCGA dataset of 1,005 patients is 0.108.



(A) FARP1 (B) FARP2

FIGURE 7.1: **Distribution of INDELs in FARP1 and FARP2.** Histograms showing the distribution of heterozygous coding cancer-specific INDELs in (A) FARP1 across 52 patients hit by INDELs in this gene and (B) FARP2 across 74 patients hit by INDELs in this gene.

TABLE 7.5: **FARP1 cancer-specific INDEL counts by tumour type.** For each tumour type, the cancer-specific heterozygous coding INDEL counts over all 1005 patients in the TCGA dataset have been tabulated for FARP1, with the mean number of INDELs per patient for each dataset (calculated to 3dp).

Disease	Number of INDELs	Mean number of INDELs per patient
BRCA	1	0.009
COAD	1	0.100
GBM	84	0.404
KIRC	2	0.011
OV	19	0.253
STAD	1	0.053
UCEC	1	0.026
Whole dataset	**109**	**0.108**

TABLE 7.6: **FARP2 cancer-specific INDEL counts by tumour type.** For each tumour type, the cancer-specific heterozygous coding INDEL counts over all 1005 patients in the TCGA dataset have been tabulated for FARP2, with the mean number of INDELs per patient for each dataset (calculated to 3dp).

Disease	Number of INDELs	Mean number of INDELs per patient
BRCA	6	0.055
COAD	2	0.200
GBM	90	0.433
HNSC	2	0.024
KIRC	4	0.023
LUAD	1	0.038
LUSC	1	0.019
OV	41	0.547
PRAD	3	0.077
STAD	3	0.158
UCEC	4	0.105
Whole dataset	**157**	**0.156**

OV generally has the highest rate of cancer-specific INDELs per patient in FARP2 with a mean number of 0.547 INDELs per patient (over all OV patients in the TCGA dataset), followed by GBM with the second highest rate of INDELs per patient with a mean number of 0.433 INDELs per patient (Table 7.6). The average number of FARP2 INDELs per patient for the TCGA dataset of 1,005 patients is 0.156, which is higher than that for FARP1.

Table 7.7 shows the INDELs in FARP1 split by their variant consequence type. Frameshift mutations are the most common with 56 of the 109 INDELs (51%) being reported as frameshift.

Similarly, in Table 7.8 frameshift is the most common type of INDEL seen in FARP2, reported at a frequency of 52% (82 out of 157). This is very similar to the proportion of frameshifts observed in FARP1.

TABLE 7.7: **FARP1 cancer-specific INDEL counts by variant consequence.** For each coding variant consequence, the cancer-specific heterozygous coding INDEL counts over all 1005 patients in the TCGA dataset have been tabulated for FARP1.

Consequence	Number of INDELs
CODON CHANGE PLUS CODON DELETION	3
CODON CHANGE PLUS CODON INSERTION	12
CODON DELETION	1
CODON INSERTION	10
FRAMESHIFT	56
SPLICE SITE ACCEPTOR	1
SPLICE SITE DONOR	3
UTR 3 PRIME	3
UTR 5 PRIME	20
Total	**109**

TABLE 7.8: **FARP2 cancer-specific INDEL counts by variant consequence.** For each coding variant consequence, the cancer-specific heterozygous coding INDEL counts over all 1005 patients in the TCGA dataset have been tabulated for FARP2.

Consequence	Number of INDELs
CODON CHANGE PLUS CODON DELETION	2
CODON CHANGE PLUS CODON INSERTION	21
CODON DELETION	2
CODON INSERTION	18
FRAME SHIFT	82
SPLICE SITE ACCEPTOR	5
SPLICE SITE DONOR	3
UTR 5 PRIME	24
Total	**157**

7.2.3.1 Clustering of FARP INDELs within patients

To examine where in the FARP patients INDELs were occurring, the degree of clustering was measured over all 52 patients containing FARP1 INDELs and all 74 patients containing FARP2 INDELs. The frequency of distances between INDELs in patients containing more than one INDEL was plotted in Figure 7.2. This histogram shows that most INDELs tend to be called within 25bp of each other, indicative of clustering of

INDELs within patients. However this plot has been truncated at 50bp along the x-axis so does not show INDEL distances exceeding 50bp, of which there are some cases. On the whole though, most INDELs can be located in clusters in the range of 0-10bp. It is possible that these INDELs reportedly being called close together are actually the same event mistaken for multiple INDELs, indicating the presence of a larger scale event taking place.

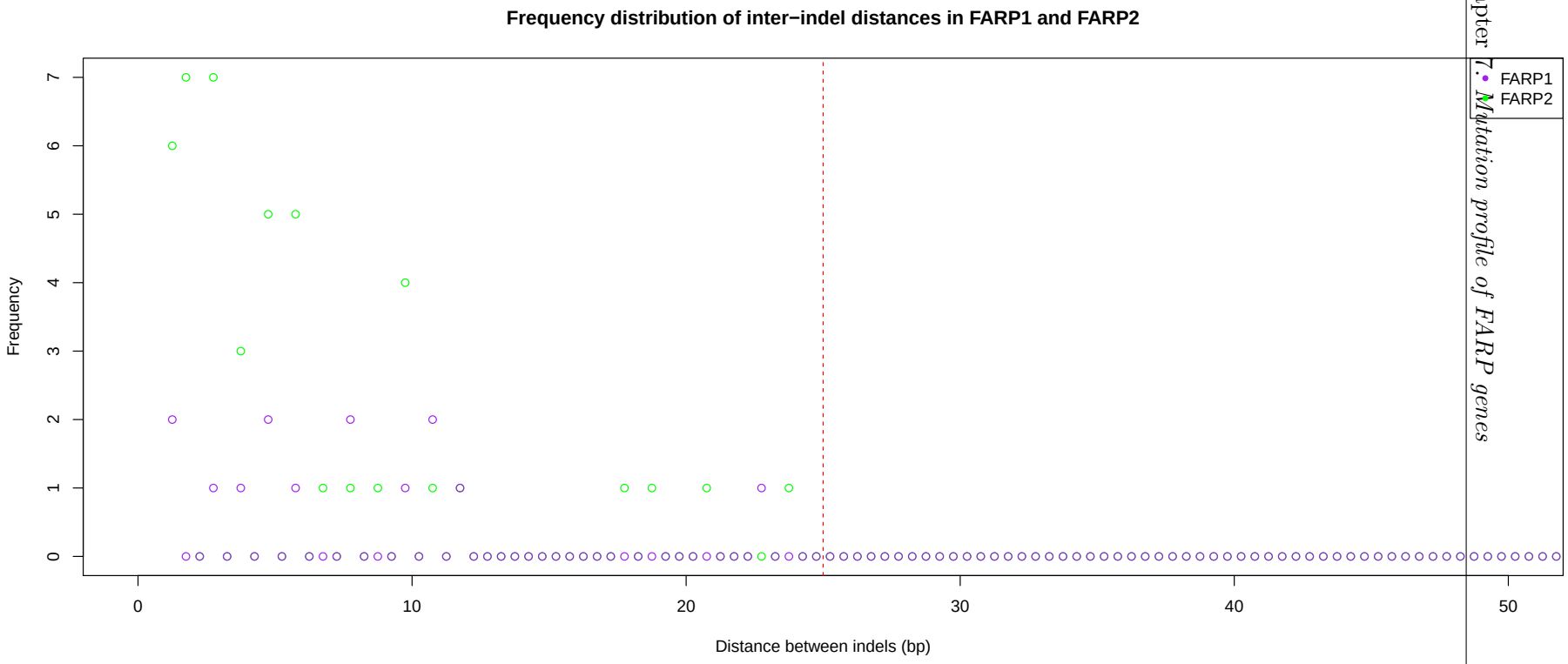


FIGURE 7.2: **Inter-indel distances in FARP1 and FARP2.** Distances between heterozygous cancer-specific INDELs in patients with FARP1 and FARP2 INDELs plotted as a frequency histogram. The frequency along the y-axis is the frequency of inter-indel distances among all 52 FARP1 patients (purple) and all 74 FARP2 patients (green). The x-axis represents the distance between two INDELs in a patient. Each patient may contain more than two INDELs in either FARP1 or FARP2 so in these cases multiple inter-indel distances will be recorded for a single patient, therefore the y-axis frequency does not necessarily represent the frequency of patients. The x-axis has been truncated at 50bp to focus on clusters of INDELs within patients. A red vertical dotted line has been drawn at a distance of 25bp, to show that INDELs tend to occur within 25bp of each other in clusters.

TABLE 7.9: **Insertions and deletions in FARP genes.** Heterozygous cancer-specific INDEL counts for FARP1 and FARP2, split into insertion and deletion sub-categories. Insertions have been split further into long insertions (≥8nt) and short insertions (<8nt).

Type of INDEL	FARP1	FARP2
Insertions	98	119
Long insertions	*94*	*99*
Short insertions	*4*	*20*
Deletions	11	38
Total	**109**	**157**

7.2.3.2 Long insertions (≥8nt)

To further investigate these INDELs, in Table 7.9 the INDELs in each FARP gene were split into insertion and deletion categories, to show that there is a much higher rate of called insertions compared to deletions. In FARP1, 90% of the INDELs are insertions, and in FARP2, 76% of the INDELs are insertions rather than deletions. Micro-insertion mutations have been sub-categorised further into short and long insertions using the arbitrary threshold of ≥8nt to define long insertions. Based on this cut-off, Table 7.9 shows that there is an abundance of long insertions compared to short insertions, with 96% of the insertions classed as long in FARP1 and 83% of the insertions in FARP2 having length ≥8nt. All subsequent analysis in this Chapter has been carried out focusing on the more abundant and easier to map (to the reference genome) long insertions (≥8nt) in FARP1 and FARP2.

7.2.3.3 GBM and OV patients enriched with long (≥8nt) insertions

In Figure 7.3, the 94 long insertions in FARP1 and 99 long insertions in FARP2 have been used to calculate the proportions of patients in each tumour type harbouring long insertions. The figure shows that OV and GBM patients are enriched with FARP long insertions. This is consistent with the results obtained for all INDELs in these patients,

which saw FARP1 and FARP2 to have higher mutation frequencies in both GBM and OV patients.

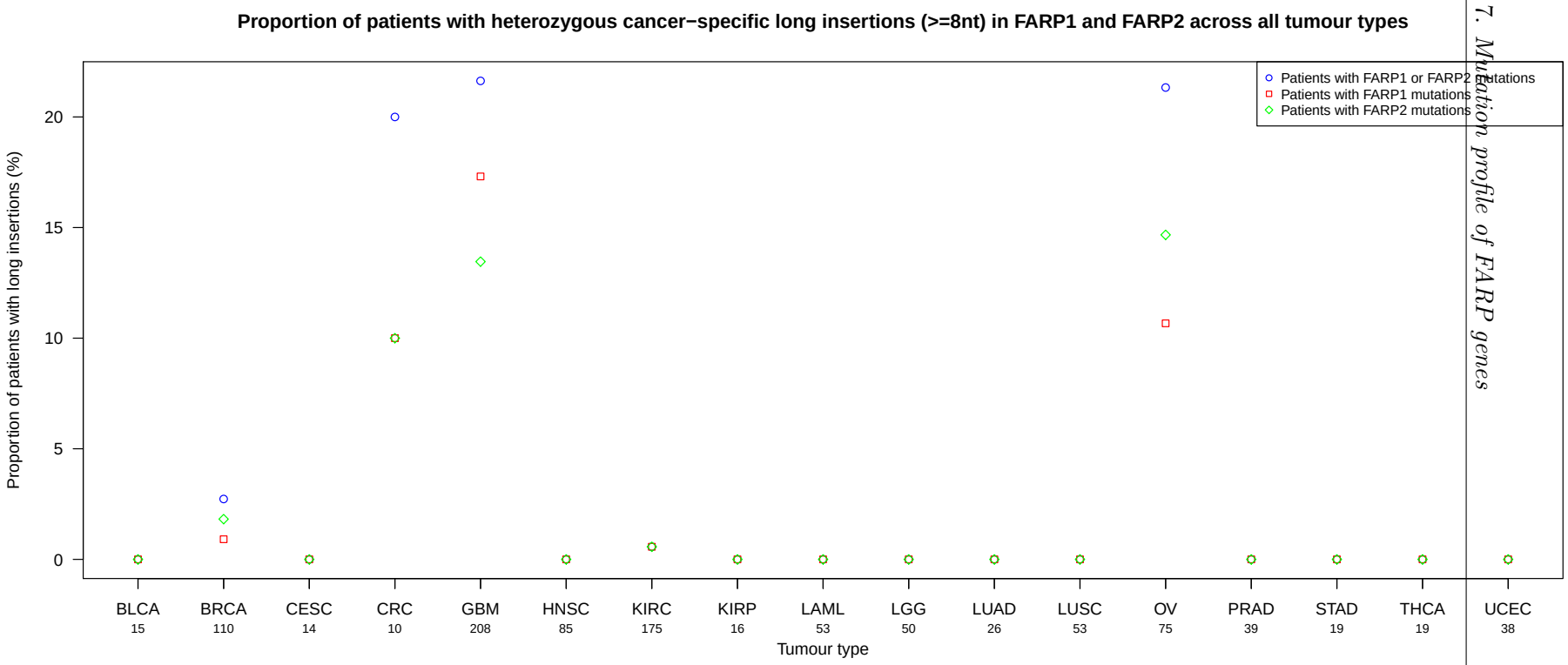


FIGURE 7.3: **Proportion of patients with long insertions (≥8nt) in FARP1 and FARP2.** For each tumour type, this plot shows the proportion of patients with coding INDELs in FARP1 (red), FARP2 (green) and either FARP1 or FARP2 (blue) out of the total number of patients for that tumour type.

7.2.3.4 Mapping long (≥8nt) insertions to the reference genome

The long insertions were aligned to the FARP loci reference genomes, in order to further ascertain the mechanism taking place in these two genes. Of the 94 long insertions in FARP1, only 27 were able to find an exact match to the reference, and in FARP2 only 30 of the 99 long insertions were able to be matched to the reference genome. Some of the unique insertions were found to map to multiple locations in the reference, so in these cases only the closest most realistic map position was considered. The results were plotted for FARP1 and FARP2 in Figure 7.4 and Figure 7.5 respectively. These INDELs are over 21 patients in FARP1 (16 GBM, 4 OV and 1 KIRC) and 20 patients in FARP2 (15 GBM, 4 OV and 1 KIRC), meaning that some patients harboured multiple INDELs within the same FARP gene. 5 of these patients had mapped long insertions in both FARP1 and FARP2 (3 GBM, 1 OV and 1 KIRC).

Figures 7.4 and 7.5 show that the positions of the called long insertions are distal to the position in the reference to which they have been mapped. This suggests that these mutations may not be micro-insertions, in which case it would be expected that the called insertions would be mapped to the same positions in the reference genome, but rather a much larger genomic rearrangement event occurring. However, since most of the called insertions have been mapped back to the FARP locus, this also suggests that the mutational process is local.

Although none of these long insertions are recurrent (occurring at the same position in multiple patients), these mapped insertions show that there is clustering of called insertion positions in the cancer sample within and among patients. For example, in Figure 7.5 there is a cluster of three red dots at position 242407653-242407682 that represent three separate insertions within the same patient. However other clusters observed are not present within a single patient and are spread over several patients.

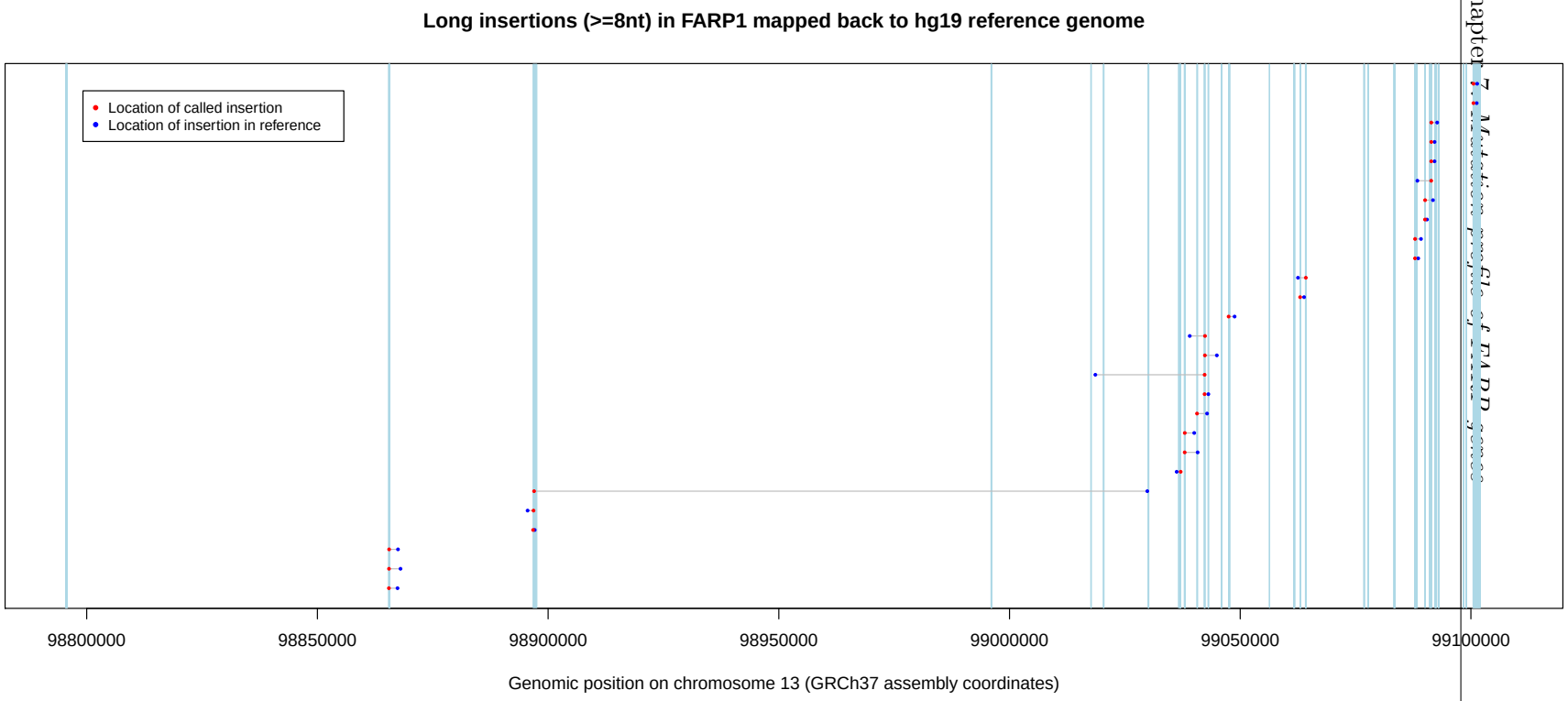


FIGURE 7.4: **Long insertions (≥8nt) in FARP1 mapped back to reference genome.** The long insertion sequences called in FARP1 have been plotted on this graph in red at the position on chromosome 13 where they have been called. In blue is the position that these sequences have been mapped back to in the reference. If a unique insertion has been mapped to several positions in the reference, only the closest has been shown on the plot. Light blue vertical segments represent exonic regions of the gene (over all transcripts for FARP1). Red points should only be found in these exonic regions since only coding INDELs have been used in this analysis, however the reference covers the whole genome so it is possible that called insertions are mapped back to intronic regions which would explain why blue points are seen outside the exonic regions. All insertions have been mapped to the forward (+) strand of the reference. *R code used to generate plot in Appendix D.*

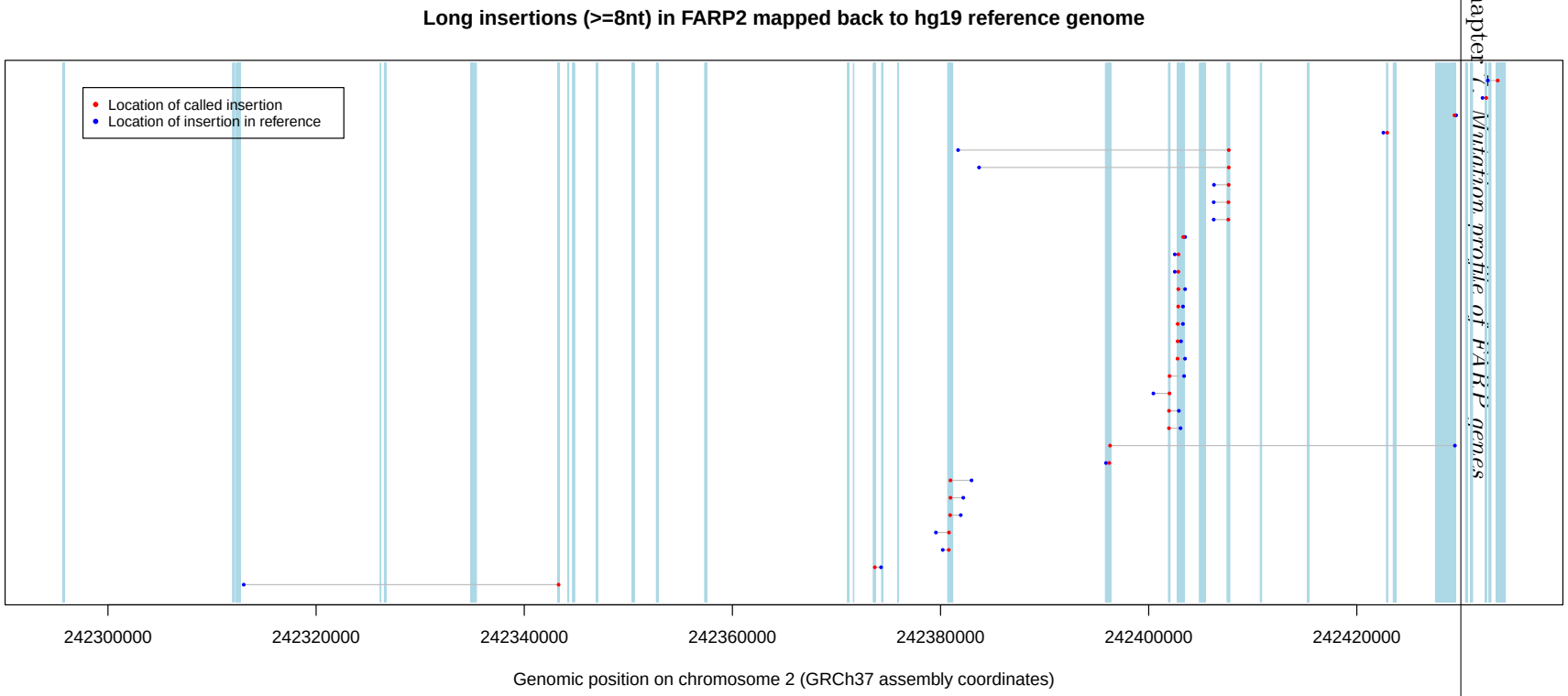


FIGURE 7.5: **Long insertions (≥8nt) in FARP2 mapped back to reference genome.** The long insertion sequences called in FARP2 have been plotted on this graph in red at the position on chromosome 2 where they have been called. In blue is the position that these sequences have been mapped back to in the reference. If a unique insertion has been mapped to several positions in the reference, only the closest has been shown on the plot. Light blue vertical segments represent exonic regions of the gene (over all transcripts for FARP2). Red points should only be found in these exonic regions since only coding INDELs have been used in this analysis, however the reference covers the whole genome so it is possible that called insertions are mapped back to intronic regions which would explain why blue points are seen outside the exonic regions. All insertions have been mapped to the forward (+) strand. *R code used to generate plot in Appendix D.*

Figure 7.6 shows the reads in the BAM file for a single GBM patient with a reported long insertion at position 98896836 in FARP1 aligned to the reference. It can be seen that an inserted sequence has been called at this position (denoted by asterisks in the reference). However what seems to actually be happening is that the reported "insertion" has come from elsewhere in the gene due to a large deletion event, hence why the reference at this point has reported the new sequence as an insertion.

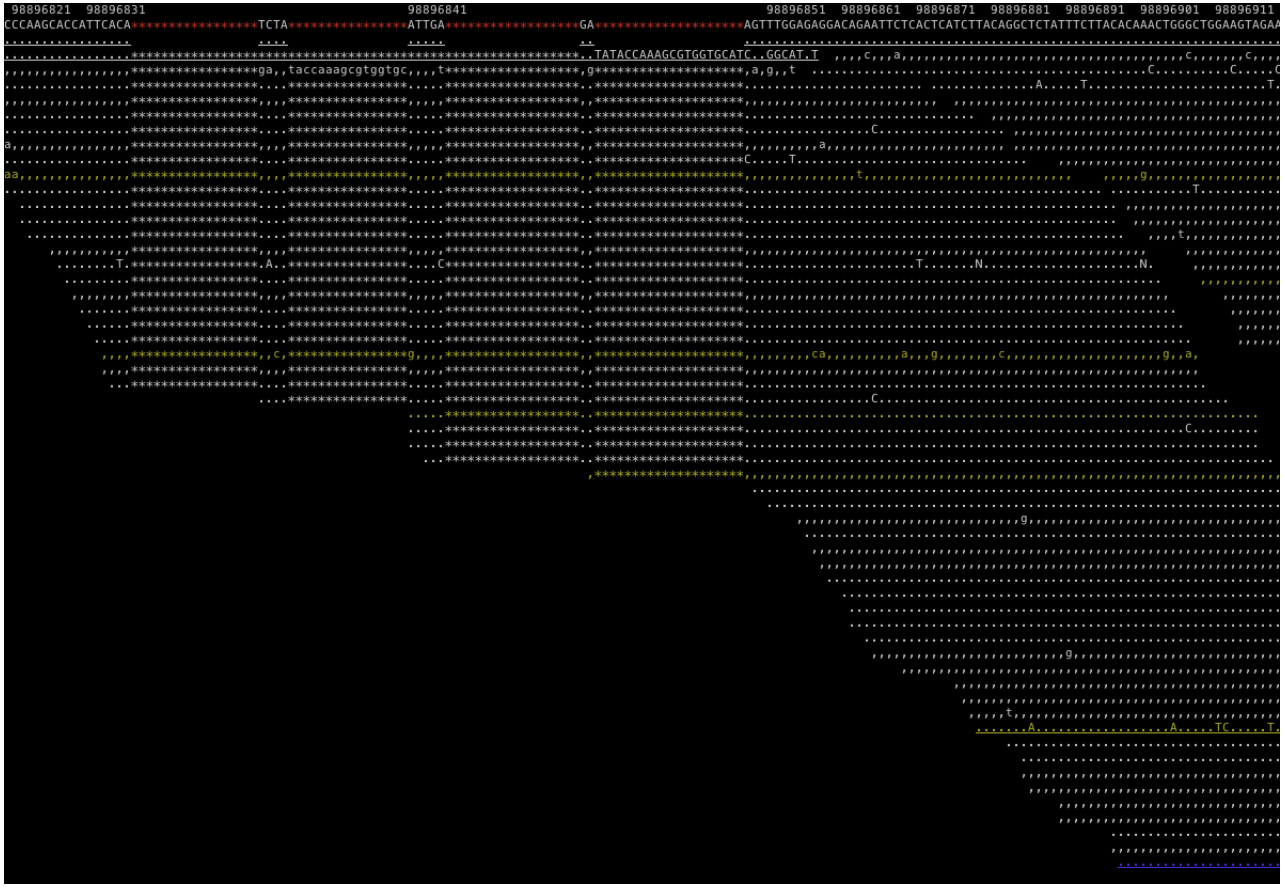


FIGURE 7.6: **Alignment of reads from GBM patient at position of called long insertion in FARP1.** BAM reads viewed through SAMtools text alignment viewer (tview) for a GBM patient with a reported long insertion in FARP1. The top line shows the genome coordinates. The second line is the hg19 reference sequence that has been provided to Samtools tview and the third line is the consensus sequence determined from the aligned reads. Uppercase bases indicate a match to the forward strand and lowercase bases letters indicates that reads match the reverse strand. Similarly a "." indicates a match to the reference on the forward strand and a "," is a match to the reverse strand. "*" in the reference indicates an insertion event. In this GBM patient a long insertion (AGATATACCAAAGCGTGG) has been reported at position 98896836.

7.2.4 Potential mutation spectra scenarios

Figure 7.7 shows the suspected mutation scenarios that could be taking place in FARP genes. It has been concluded that an insertion as reported by GATK is not taking place, because reads do not support the capture of two edges ("ab" and "bA" as shown in Figure 7.7) which would be expected to be seen in the case of a long insertion. Instead only one edge is captured.

Since long insertions are known not to be occurring in these genes, other mechanisms must be explored. There are three other possibilities (as shown by Figure 7.7):

- Large-scale segmental deletion - one edge ("ab") would be detected.
- Tandem duplication - change in order of sequence ("BA").
- Inversion - opposite orientation would be seen (reverse strand).

The relative orientations of the edges of the called insertion sequences have been investigated and they are consistent with a deletion ("ab"). These events cannot be tandem duplications since the order of the sequence has not been changed as would be expected to happen with tandem duplications. Inversions were also ruled out since all called insertion sequences were mapped to the forward strand and if an inversion was taking place then reverse compliments would be seen where the inserted sequences would have mapped to the reverse strand.

Chromothripsis "chromosome shattering" was also hypothesised to be a potential cause of this mutation spectra. This is a process in which one or a few chromosomes in a cancer cell bear dozens to hundreds of clustered rearrangements [Forment et al., 2012]. However this would also involve different orientations (+ and - strands) being observed which is not seen, and would not be a local event.

In some cases in Figure 7.4 and Figure 7.5, the insertion sequence has been mapped downstream in the reference (blue dot to the left of red dot), and in other cases the insertion sequence has been mapped upstream (blue dot to the right). The orientation

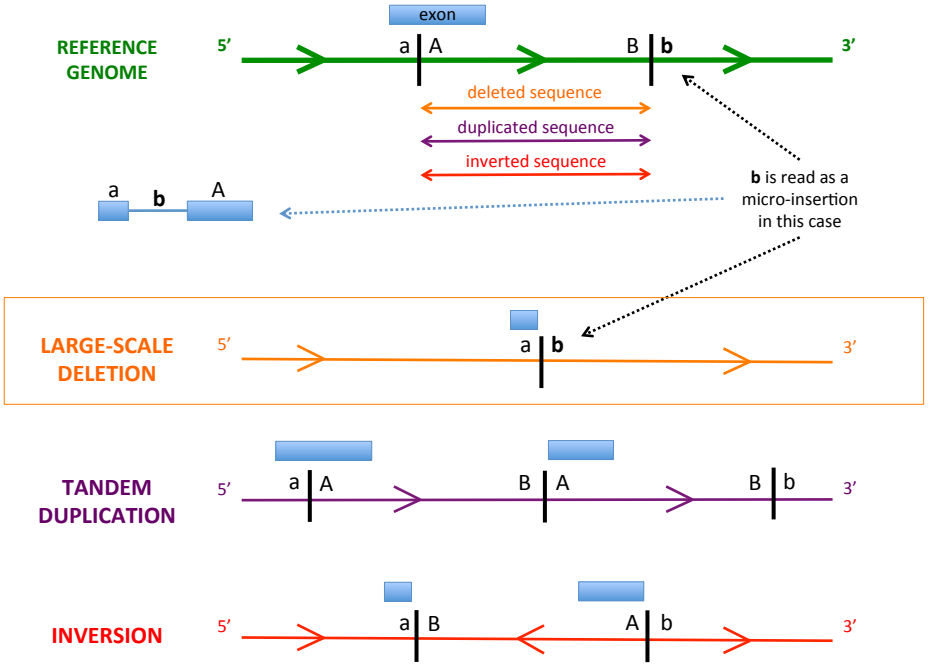


FIGURE 7.7: **Possible mutational events occurring in FARP genes.** This figure shows the possible mutation scenarios that could be taking place in FARP genes: large-scale deletion, tandem duplication and inversion. The disrupted exon is located in the reference at "Aa". The "b" from the reference genome has been reported as an insertion in these genes, however it is suspected that large scale-deletions are actually taking place so that the sequence in the cancer sample reads "ab". If the disrupted exon read "Bb" in the reference sequence then "a" would be reported as the reference sequence.

depends on the type of event that has occurred. If a deletion is occurring as is suspected, then a red dot to the left of the blue dot would correspond with Figure 7.7. The other way around (blue dot to the left of the red dot) would suggest that an exon at "Bb" had been disrupted with "a" read as the insertion sequence.

7.3 Discussion

7.3.1 Unusual mutation spectrum

An unusual mutational profile was observed in the FARP1 and FARP2 genes in the TCGA dataset: a mutator phenotype exhibiting a high rate of called INDELs. On closer inspection, this phenomenon was discovered to be specific to cancer samples and not detected in the control samples, suggesting that these mutations were implicated in the progression of cancer either as driver mutations or as passengers highlighting a mutational mechanism underlying the progression of these cancers.

7.3.2 Miscalled large-scale deletions

The called INDELs were found to be mostly small insertions. Modelling of these insertions showed that all reads around the called insertions within these genes were co-orientated and mapped to the forward strand at distal positions in the reference genome. This lead to the conclusion that the called insertions were not actually micro-INDELs at all, but instead a local mechanism consisting of larger segmental deletions occurring within and limited to the FARP locus.

7.3.3 Measures of selection

This is a rare and interesting phenomenon, however it is not known whether selection is occurring as opposed to mutation. Unfortunately there are no evolutionary models in place in this project to look at any mutation type other than point mutations, as there would be no background model and hence no null hypothesis. Therefore, it has not been possible to determine whether these deletions are drivers of cancer or inconsequential passenger events.

7.3.4 Validation

It would be useful to validate such findings. One way in which this can be done is by using coverage data available for the TCGA dataset. Reduced coverage at FARP loci would confirm deletions as the mode of mutation in these genes.

7.3.5 Genome-wide phenomenon in GBM

It is possible that this mechanism is not just restricted to the FARP loci. To decipher whether this is a gene-specific or genome-wide phenomenon, the rate of micro-insertions called in the TCGA dataset could be analysed over the whole exome. GBM and OV cancers were the cancer types exhibiting the highest enrichment of large-scale deletions in FARP genes. Since FARP1 and FARP2 are known to be involved in function of the brain, a genome-wide analysis in GBM would be the most interesting to pursue. This would also support the results from Chapter 3 in which a subset of GBM patients were found to exhibit a very high number of INDELs. It is possible in light of the results from this chapter that these INDELs are also representative of a larger scale rearrangement that has been miscalled. These particular INDELs were called over multiple genes across the exome, supporting the hypothesis that this is a genome-wide phenomenon. It is possible that in these 16 GBM patients with high INDEL rates, the driver mutations are located within FARP genes, however 20% of the 208 patients (42) in the TCGA dataset contain large-scale deletions in FARP genes, suggesting that the 16 GBM patients alone cannot account for this mutator phenotype, so it is highly possible that other genes are also experiencing this structural rearrangement event in GBM patients.

7.3.6 Implications of novel discovery

There are many potential events that could be causing this observed mutation phenotype in FARP genes. For example, large-scale deletions could be an artefact of necrosis.

However, this is unlikely in the TCGA dataset since samples have been filtered so that only those with less than 50% necrosis have been sequenced.

The clustering seen in Figure 7.4 and Figure 7.5 could also be an artefact of using exome sequencing (exome-seq) data rather than whole-genome sequencing (WGS) data. WGS is superior at detecting genomic rearrangements compared to exome-seq [Meyerson et al., 2010], and these results could be highlighting the limitations in using exome-seq rather than WGS.

Additionally, this chapter highlights the problems encountered with current variant calling tools that are not capable of accurately calling INDELs when other larger structural rearrangements are present in cancer genomes. A similar problem was illustrated in Chapter 3, in which SNVs were incorrectly called in the presence of INDELs.

Therefore these limitations should be considerations for all cancer studies during variant identification stages especially if exome-seq is used, since the presence of genomic rearrangements could have a detrimental affect on both the SNV and INDEL results, as has been shown both in this Chapter and in Chapter 3.

However, since this phenomenon has been detected solely in cancer exomes, it is unlikely to be an artefact of variant calling or exome sequencing. It is more plausible that the large-scale deletions observed in FARP1 and FARP2 are demonstrating a mutation spectrum defined by genomic instability, a characteristic of most cancers [Negrini et al., 2010]. Moreover, extensive genomic instability is known to be one of the main features of malignant gliomas [Milinkovic et al., 2012], in which FARP deletions have been shown to be enriched. This supports the roles of FARP1 and FARP2 as putative candidate cancer genes in GBM.

7.4 Methods

7.4.1 FARP SNVs

The filtered set of heterozygous coding cancer-specific SNVs (non-synonymous and synonymous) from the TCGA dataset used in Chapter 3 and in Chapter 4 for evolutionary analysis were used here to look specifically for SNVs occurring in FARP1 and FARP2.

For these SNV counts, the mutation occurring on the transcript with the most severe consequence was counted for each unique SNV, counting a specific mutation only once if occurring in overlapping genes.

7.4.2 FARP INDELs

To create the tabulated cancer-specific coding heterozygous INDEL mutation counts for each gene by variant consequence and tumour type, first the mySQL TCGA database (tcga_pair_exome) was mined using the command in Listing 7.1. Genotypes as determined by GATK were used to specify cancer-specific INDELs. This is a less sensitive approach than using alt read depth (0 in control and >0 in cancer) which is prone to more false-positives and cannot be used to identify heterozygous only INDELs. Only coding INDELs were selected from the database, since the consequence table was used which only contains information for coding mutations.

```
1 select gene_name, patient_id, var_site.type, chr, pos, ref, alt,
    consequence.type, disease, cancer_genotype, control_genotype,
    cancer_genotype_qual, control_genotype_qual, gene, transcript from
    sample, consequence, var_site, var_site_sample where sample.sample_id
    = var_site_sample.cancer_sample_id and consequence.var_site_id =
    var_site.var_site_id and var_site_sample.var_site_id = var_site.
    var_site_id and (gene_name = "FARP1" or gene_name = "FARP2") and
    control_genotype_qual >= 30 and cancer_genotype_qual >= 30 and
    var_site.type = "INDEL" and cancer_genotype = "0/1" and
    control_genotype = "0/0" and patient_id != "TCGA-AB-2942" and
    patient_id != "TCGA-EJ-5515"
```

LISTING 7.1: Mining mySQL database for cancer-specific heterozygous FARP INDEL mutations

Patients TCGA-AB-2942 and TCGA-EJ-5515 were not selected from the TCGA database for consistency with SNV analysis, since these patients have no cancer-specific SNVs and so were not included in the SNV analysis in Chapter 3 or the evolutionary analysis in Chapter 4.

The INDEL set was filtered to contain just the unique mutations for each gene for each patient, by recording the most severe consequence for each INDEL (Perl code in Appendix C).

INDELs were counted by variant consequence and tumour type, using the R code in Listing 7.2.

```
#read in unique list of FARP INDELs
FARP_indels<-read.table("FARP_indels_unique", header=T, sep="",quote="")
FARP_indels$count=1

#create subset of FARP1 INDELs
FARP1_indels<-subset(FARP_indels, (FARP_indels$gene_name == "FARP1"))

#create subset of FARP2 INDELs
FARP2_indels<-subset(FARP_indels, (FARP_indels$gene_name == "FARP2"))

#count number of INDELs in FARP1 by consequence type
aggregate(FARP1_indels$count, by=list(cons=FARP1_indels$cons), sum)

#count number of INDELs in FARP2 by consequence type
aggregate(FARP2_indels$count, by=list(cons=FARP2_indels$cons), sum)

#count number of INDELs in FARP1 by tumour type
aggregate(FARP1_indels$count, by=list(disease=FARP1_indels$disease), sum)

#count number of INDELs in FARP2 by tumour type
```

```
21 aggregate(FARP2_indels$count, by=list(disease=FARP2_indels$disease), sum)
```

LISTING 7.2: R code to count FARP INDELs by consequence type and tumour type

For the counts by disease type, to calculate the mean number of INDELs per patient for each tumour type and across the whole dataset of 1,005 patients, each total mutation count across all patients with that disease type was divided by the total number of patients with that disease type, for a direct comparison of mutation rates per patient between tumour types.

7.4.3 Mapping long insertions (≥8nt) to reference

The called INDEL insertion sequences were mapped to the hg19 reference sequence for each gene by using Perl to find a match for the alt allele sequence reported as an insertion in the reference sequence. Only the closest match was plotted for each unique insertion event using R code in Appendix D.

7.4.4 SAMtools tview

Samtools tview is a text alignment viewer [Li et al., 2009]. The reference genome for FARP1 was input to tview, together with the target cancer sequence BAM file of a patient known to have an insertion in FARP1 in the cancer sample. Using this function, the known position of the called insertion can be specified in order to jump straight to that part of the genome.

Chapter 8

Discussion

8.1 Concluding remarks

In this project I set out to identify cancer-specific single nucleotide variants in cancer exomes obtained from The Cancer Genome Atlas, and distinguish the important driver mutations from the inconsequential passenger mutations using methods from the field of molecular evolution which allowed the detection of signals of positive selection in cancer genomes to highlight genes enriched with driver mutations.

I also wanted to partition the data to ask how tissue of origin and mutation spectra of the cancer influences the path of cancer development and the genes that are hit by driver mutations. Additionally I planned to partition the data by functional sub-regions such as kinase domains, to ask more interesting questions such as where exactly in driver genes the driver mutations are targeted, to further understand the mechanism underlying cancer progression. It was also intended to extend this sub-region analysis to pathway analysis, concatenating a domain of interest over a pathway of interest to obtain more power in the evolutionary analysis. Greenman et al. [2007] have done this in 518 kinase genes, by estimating the selective pressures in the kinase domains of these genes compared to regions outside the kinase domains, showing that the selective pressure is higher in the kinase domains and even higher in the P-loops and activation

segments of these domains, suggesting a greater selection pressure for mutations within kinase domains and further validate their role in cancer. In this work pathway analysis was also considered although the selection pressures across pathways were not estimated. Clustering of mutations in genes involved in the JNK-pathway indicated that mutations in this pathway are involved in cancer development, and the FGF signalling pathway showed the highest enrichment for non-synonymous mutations. Both these pathways would therefore be good candidates for pathway analysis in PAML.

I have successfully called cancer-specific variants in 1,005 TCGA patients, and performed gene-based evolutionary analysis on the whole dataset consisting of 17 different tumour types.

Several key papers relating to this work were published during the course of this project, including studies using Pan-cancer data from TCGA. These included the Lawrence et al. [2014] and Kandoth et al. [2013] studies. These publications helped shape the project by redirecting it, since these papers had already covered some of the ground originally aimed for in this project by addressing the question of tissue of origin [Lawrence et al., 2014] and mutation spectra [Kandoth et al., 2013] based analysis.

The much larger Lawrence et al. [2014] dataset was used to complement the fully processed TCGA dataset. Mutation data from Lawrence et al. [2014] was stratified by tumour type, as was done in Lawrence et al. [2014], however gene-based evolutionary analysis was then performed, and the results from the two different methodologies (PAML and MutSig) were compared. As the methods in this project were distinct from the MutSig analysis in Lawrence et al. [2014], there was merit in performing the complementary analysis.

The Lawrence et al. [2014] data was also used to stratify the data by mutation spectra, which had not been done by Lawrence et al. [2014], using the six classes of single nucleotide mutations to classify six distinct mutational signature groups. Mutation spectra was addressed in the Kandoth et al. [2013] study, where they have also used six mutation categories. However again there is merit in performing this analysis, since

Lawrence et al. [2014] have not considered mutation spectra so it is building on the work in that analysis, and Kandoth et al. [2013] have used different methods to identify significantly mutated genes.

Although Yang et al. [2003] and others have used codon models to look at cancer evolution (analysing many patients over a single gene), and omega ratios have previously been used in a cancer setting, what is new here is the systematic analysis of all genes in the same set of patients. Innovations exist in the way such analysis has been implemented, by editing variants onto a reference, adapting PAML to deal with missing data and handle stop codons, as well as the huge scale of the analysis.

8.1.1 Comparison of called variants between TCGA and Lawrence datasets

701 patient samples containing coding variants were shared between both the TCGA and Lawrence datasets, permitting a direct comparison of their called variants by the two different pipelines. I found results to be broadly consistent except for a subset of GBM patients where SSNV calls were considerably more frequent in my analysis than in the Lawrence analysis. This has been traced to an issue with INDELs, in which it was suspected that a higher rate of INDELs in a subset of GBM patients was causing a higher rate of miss-called SSNVs, likely to be due to the difficulty in calling SSNVs around INDELs using current variant identification tools such as GATK. This discrepancy represents the limitations of SSNV calling pipelines, but shows that Lawrence is the more conservative approach.

8.1.2 Stratification of patients by tissue of origin and mutation spectra

I set out to address how tissue of origin and mutation spectra affect the genes hit by driver mutations in tumour samples, and which factor has more impact on the path

of cancer development, to further the understanding of the mechanisms responsible for the development of different cancers.

It has been shown in this project that the tissue of origin does not necessarily correlate with the mutation spectrum of a particular cancer, and so it can be concluded in these datasets that the mutational spectrum is not necessarily dependent on the tissue of origin, therefore both the varying mutation spectra (observed over TCGA and Lawrence datasets) and the tissue of origin were controlled for separately in the evolutionary analyses of the Lawrence SSNVs. This was done by stratifying the data by tissue of origin and mutation spectra before detecting selection at the gene level.

The results were compared and contrasted within and between the two types of stratification methods to understand how each relates to the path of selection, and to ask if the path of cancer development is dependent on mutation spectra. There are clear contrasts between the tissue of origin and mutation spectra analyses, with different patterns of positive selection and different genes reaching significance in the two types of analysis. Within the mutation spectra analysis, the genes targeted by driver mutations differ between the six different mutation profiles, showing that certain genes are more prone to being hit by certain types of mutation. It is therefore hypothesised that mutation spectra is more important than tissue of origin in determining where the selection is acting, however experimental validation is needed in order to answer this question.

However, it was also shown that some specific mutation spectra are tissue-restricted. For example, lung cancers caused by smoking are known to have a specific mutation spectrum with an increased rate of C→A mutations as was shown in Chapter 3. Melanoma caused by UV is another cancer type induced by environmental exposures known to have a specific mutation spectrum of extremely high rates of C→T mutations. However, within these cancer types, variation in mutation spectra was also observed suggesting that the mutation spectrum of a cancer as well as the tissue of origin should be used to diagnose and treat individual cancers. Other known specific mutational spectra induced by endogenous mutator mechanisms, such as loss of mismatch repair or loss of

DNA polymerase proofreading activity, are common to multiple cancer types, further confirming the importance of mutation spectra in the characterisation of a cancer.

8.1.3 Significantly mutated novel candidate cancer genes

Evolutionary analysis in PAML has identified two novel candidate cancer genes not previously implicated in cancer, that were significantly mutated in the Lawrence dataset using codeml but were not picked up in the Lawrence et al. [2014] study.

In the tissue of origin analysis in Chapter 5, DNMT1 was found to be significantly mutated in colorectal cancer (CRC). DNA methyltransferase I (DNMT1) is the primary regulator of DNA methylation patterns in mammalian cells [Song et al., 2015]. Altered methylation of DNA was among the first of the genetic aberrations identified in CRC [Goel and Boland, 2012]. DNA methylation is the most prominent type of epigenetic change. Epigenetic changes broadly encompass all alterations in the regulation of gene expression that do not involve a change in the DNA sequence, occurring through modified interactions between the regulatory portions of DNA or messenger RNAs (mRNAs) [Goel and Boland, 2012]. Epigenetic change usually occurs through changes in the promoters of genes, modification in the stability of transcripts or alterations in the splicing of transcripts. However, genetic variants in a gene or regulatory element distant may also cause epigenetic changes by modifying DNA methylation for example. DNA Methylation is mediated by DNA methyltransferases (DNMTs) such as DNMT1, that act by catalysing the covalent addition of a methyl group to the 5' carbon of cytosine creating 5-methyl-cytosine in GC-rich promoter (CpG islands) regions. This mechanism can completely silence gene expression by preventing transcription factors from interacting with the DNA. In this analysis, genetic SNVs have been detected in DNMT1. It is therefore predicted that the putative driver mutations detected in this gene are activating mutations that cause the DNMT1 to hypermethylate tumour suppressor genes at their CG-rich promoter regions, diminishing the expression of the tumour suppressor genes, leading to the development of cancer.

In the mutation spectra analysis in Chapter 6, DNA polymerase theta (POLQ) was found to be significantly mutated in Signature 3. This enzyme is a member of the DNA polymerase A-family, which function by synthesising DNA in genome replication and protecting the cell against DNA damage [Lange et al., 2011]. Another replicative DNA polymerase, POLE, has recently been identified as a known cancer gene, which has been found to harbour somatic mutations in sporadic colorectal and endometrial cancers, as well as germline mutations predisposing to colorectal adenomas and carcinomas [Heitzer and Tomlinson, 2014]. However, POLQ has not yet been identified as a cancer gene. The role of this gene in cancer has recently been investigated in [Ceccaldi et al., 2015] and Mateos-Gomez et al. [2015], in which POLQ was identified as a key factor in the error-prone alternative non-homologous end-joining (alt-NHEJ) repair pathway, acting by suppressing homologous recombination (HR). Furthermore, it was found that the inhibition of POLQ in cancers with defective HR, through the absence of homology-directed repair genes such as BRCA1 or BRCA2, had a synthetic lethal effect on cell survival, presenting POLQ as a promising therapeutic target in HR-deficient tumours. In the context of this research, it is suspected that the SNVs identified in my analysis are activating mutations causing the up-regulation of POLQ which in turn promotes alt-NHEJ and suppresses HR, causing a higher rate of translocation and genomic instability leading to the mutation spectrum characteristic of many cancers. In HR-deficient tumours, such as epithelial ovarian cancers with mutations in genes involved in HR, activating mutations in POLQ would be expected to further promote the use of the compensatory alt-NHEJ repair pathway. Four mutations in POLQ were mapped to the polymerase domain, which is required for the function of alt-NHEJ, and one mutation was located in the helicase ATP-binding domain, a region found to be required for the inhibition of HR as well as the disruption of RAD51 foci which mediate HR. This further supports the POLQ mutations identified in this PAML analysis as activating candidate drivers of cancer. Interestingly, POLQ was not identified in any of the tissue type analyses, highlighting the merit of stratifying data by mutation spectra to identify candidate cancer genes that would otherwise be missed

when only tissue of origin is considered. POLQ has been identified as significantly mutated in just one distinct mutational signature, characterised by a mutational pattern typical of HR-deficient tumours, further confirming the role of POLQ in suppressing HR in cancers.

8.1.4 FARP mutation profile

An interesting mutation profile characterised by a high incidence of INDELs was observed in FARP1 and FARP2 genes in the TCGA dataset. Further investigation revealed that the reported INDELs were actually large-scale deletions. Further analysis is required to discover whether this is a genome-wide phenomenon or whether it is restricted to these two genes in the human genome.

The high rate of called INDELs in FARP1 and FARP2 was more pronounced in GBM and OV patients, which is in agreement with the high rate of INDELs that was observed in a subset of GBM patients in Chapter 3. This is suggestive of this event being less specific to FARP genes, and could suggest the presence of large-scale deletions in other genes of the genome.

Glioblastoma multiforme (GBM) is a malignant brain cancer and is the highest grade glioma, with an annual incidence of five cases per 100,000 people [Patanè et al., 2013]. The thorough characterisation of this cancer type has lead to the classification of GBM into four molecular subtypes: classical, mesenchymal, proneural and neural. GBM develops after the accumulation of genomic DNA damage that often includes gene amplifications and/or deletions [Rao et al., 2010], further supporting the hypothesis that this large-scale deletion phenomenon is genome-wide rather than specific to FARP genes. The classical subtype is mostly characterized by loss of chromosome 10 and amplification of the epidermal growth factor receptor gene (EGFR). However, heterozygous deletions have been described in the NF-kB Inhibitor alpha gene (NFKBIA) in about 20% of GBM cases, so this gene may be a good target for the search of large deletions in the TCGA dataset. These deletions were found to be mutually exclusive with EGFR

amplifications, which is a frequent event in GBM [Patanè et al., 2013]. EGFR has been detected as highly significantly mutated by SNVs in my PAML analysis, suggesting that, if amplifications that cause overexpression are a common aberration in GBM, these point variants may be activating mutations. Additionally, it is suggested that a subset of GBM patients without EGFR amplifications are subject to deletions which would indicate that the deletion spectrum observed in the FARP genes may not be present in all GBM patients and specific only to a subset.

These findings highlighted the lack of power to successfully detect INDELs and large-scale genomic rearrangements in the TCGA pipeline. Whole-genome sequencing data would be required in order to accurately identify the presence of large-scale rearrangements in GBM patients.

8.1.5 Challenges and limitations

The benefit of using sequencing data from TCGA is that very large amounts of data were able to be obtained for meta-analysis. However, with this came certain technical challenges in the bioinformatic analysis of the next-generation sequencing data, all of which could have lead to systematic biases in the subsequent analyses. Throughout this project, I have attempted to remove these biases.

Data consistency was a problem encountered with the exome sequences obtained from TCGA, since not all sequences had been aligned using the same method and some were aligned to an older version of the reference genome, hg18. To remove this inconsistency, in the TCGA re-analysis all exomes were realigned to hg19 using a consistent pipeline.

Another challenge encountered was the varying coverage across the genome, which is a major source of false-negative cancer-specific variants, where there is insufficient coverage to confidently call a true variant in the tumour sample. This is not well dealt with in other similar studies, however it has been quantified and accounted for in this analysis by annotating sites of low coverage (<10X) as missing data before evolutionary analysis, so that selection is not underestimated.

The scale of data was also a huge challenge, with over 40Tb of exome data downloaded from TCGA (∼20Gb per patient), requiring a large amount of time and computational resources. Working with this large quantity of data was a logistical challenge and had the potential to make the analysis very time consuming and difficult to handle. The realignment and variant calling processing part of the pipeline alone took approximately eight-ten days per patient, the time-limiting step attributed to the use of Stampy to re-align exomes, taking up substantial CPU time to process. Therefore the need for efficient pipelining was non-trivial, and was successfully put in place to manage this computationally.

A problem encountered generally in cancer studies using primary tumour samples is that of cellular heterogeneity. This refers to the presence of multiple different sub-clonal populations within a tumour, so not all cells within a tumour population will contain the same ancestral mutations. This can be a problem in exome sequencing, since sequencing DNA from cells that have different cancer-specific mutations reduces the power to detect these mutations, especially heterozygous mutations that are expected to be at a frequency of about 50% in the tumour population, whereas homozygous mutations should have a frequency of 100%. Cell lines can be used to avoid this problem, and are available from TCGA, however these also have their disadvantages when evolutionary models are used. Cell lines are grown in culture so the advantage is that there is less heterogeneity. However, although these cells carry the genomic mutations from their source tumour samples, they also gain additional mutations during the course of cell line development and passage [Chang et al., 2011], which will confound any evolutionary analysis when studying the effects of selection. Also cell lines that grow out of a tumour do not just have the possibility of additional mutations, they may have represented a very minor sub-population of the parent tumour, and may not be representative of the driver mutations that have caused the cancer. This is the justification for using primary tumour samples. It has however been shown by Neve et al. [2006] and Forbes et al. [2010] that there are no significant differences in the spectrum of genomic mutations between cell lines and primary tumours. Another option to help address this issue

would be to sample tumours over time so that new mutations appearing in a clonal population could be reported, however this was not addressed in this analysis.

A particular weakness of the work in the TCGA data processing pipeline is the miscalling of INDELs in FARP1 and FARP2 genes, suggesting that Genome Analysis Toolkit (used to call both SNVs and INDELs in the TCGA dataset) may also be calling INDELs incorrectly across the whole exome and potentially representing larger scale mutations. This technical challenge has not affected the evolutionary results since only SNVs were used for this, however it does impact on the SNVs that are called, since it is known that SNVs are miscalled near INDELs, as has been shown in a subset of GBM patients in Chapter 3.

Despite the PAML evolutionary models used in my TCGA analysis not explicitly modelling the clustering or recurrence of mutations, in general I observed good sensitivity for detecting genes harbouring clustered mutations. For example, PIK3CA was regularly identified and has highly clustered mutation patterns. However in the most extreme cases such as BRAF where a single codon is specifically mutated in many cancers, including colorectal cancer, this gene was not detected in the PAML analysis whereas it was identified in the MutSig analysis in Lawrence et al. [2014]. This appears to a limitation of our approach where the evolutionary signature is too homogeneous. The MutSig analysis seems better at dealing with this caveat, likely to be due to the incorporation of a MutSigCL statistical test in the Lawrence et al. [2014] study which accounts for the clustering of mutations in genes.

A high false-positive rate has been observed in implausible cancer genes, such as those encoding olfactory receptors, in the PAML analysis. It seems that the PAML approach is more prone to picking up likely false-positives, though Lawrence also shows some such tendencies. This is a fundamental problem with cancer genome studies, in that as the sample size increases the number of false-positives detected also rises [Lawrence et al., 2013]. Despite many of these suspected false-positives genes encoding large proteins such as the muscle protein titin (TTN) and membrane-associated mucin (MUC16), the prominence of these genes in cancer studies is not simply the consequence of the

length of their coding regions since statistical tests such as the one used in this PAML analysis already account for the large target size. Instead it has been found that the source of this problem stems largely from heterogeneity of mutational processes in cancer. This refers to three types heterogeneity: the heterogeneity of mutation rates across patients with a given cancer type as well as across patients within a cancer type; the heterogeneity in the mutational spectrum of tumours (using 96 possible mutations incorporating the single nucleotide mutation and the sequence context); and the most important type, regional heterogeneity across the genome with differences exceeding fivefold [Lawrence et al., 2013]. The latter heterogeneity type is known to be in part caused by DNA replication time, with late-replicating regions exhibiting much higher mutation rates, and gene expression level with low expression correlating with a higher mutation rate. Both factors could explain the high frequency of somatic mutations in large genes and olfactory receptor genes, which are known to have low expression and be late-replicating, and hence explain why these genes are often presented as putative cancer-associated genes [Lawrence et al., 2013]. Lawrence et al. [2014] have attempted to control for this in their methods by incorporating mutational heterogeneity into the analyses through the use of MutSigCV to eliminate false-positives Lawrence et al. [2013]. A possible fix is needed in the PAML analysis to also account for this.

Having identified limitations in my own analysis, caveats in the Lawrence analysis were also observed. For example, the varying depth of coverage across the genome was not quantified and accounted for in the Lawrence et al. [2014] analysis, as it was in my TCGA analysis by annotating regions of low coverage as missing data in the exome sequence alignments prior to driver mutation detection in PAML. This is expected to result in a higher rate of both false-positives and false-negatives where coverage is too low to confidently call a variant in the control and cancer sequences respectively. Lawrence et al. [2014] have also not used a uniform dataset with all exome sequences re-aligned to the hg19 reference genome, as has been done in the TCGA dataset. Instead they have used the UCSC Liftover tool to convert coordinates of each mutation to build hg19 for tumour types originally aligned to build hg18. Although this is less time consuming than re-aligning to hg19, the results will be less accurate and consistent.

Both of these confounding factors will have introduced systematic biases into their analysis, but have been taken into account and corrected for in the TCGA pipeline resulting in an improved and more rigorous methodology.

Despite these caveats, Lawrence et al. [2014] provided a very large dataset to complement the smaller TCGA dataset, which added power to the previous preliminary evolutionary analysis, and allowed for a direct comparison of different driver gene detection methods.

8.2 Future research

8.2.1 Validation of results

Evolutionary analysis in PAML has uncovered putative candidate cancer genes found to be under positive selection in both the tissue of origin stratified sub-analysis and the mutation spectra stratified sub-analysis. These results require both computational and experimental validation to further provide evidence of their role in tumourigenesis.

8.2.1.1 Computational validation of results

If a loss of function is suspected to have taken place in a candidate cancer gene, the result could be validated by looking for evidence of loss of heterozygosity (LOH) at the locus. This can be done by examining the coverage data, however this is not available for the Lawrence dataset. Additionally cancer-based transcriptome sequencing (RNA-seq) data can be used to look for reduced expression in that gene which would support the evidence for a loss of function. Another validation process would be to look for germline mutations in the relevant datasets to see if they correlate with detected somatic mutations, for example by using population exome cohort data containing germline variants present in colorectal cancer to support findings (Dunlop and Farrington, unpublished data). For example a heterozygous germline mutation in a suspected tumour suppressor

gene thought to have undergone a loss of function mutation would support the tumour suppressor hypothesis.

Computational validation using other types of data such as array-based expression, protein expression, copy number, microsatellite instability, DNA methylation and clinical data is also available from TCGA, and would further support evidence for potential candidate genes. The Catalogue of Somatic Mutations in Cancer (COSMIC) [Forbes et al., 2010] can also be used as a resource to cross-reference the significant evolutionary results.

8.2.1.2 Experimental validation of candidate cancer genes

Experimental follow-up in the laboratory is essential in order to complement these novel discoveries and validate their role in tumourigenesis.

A classical approach is to functionally validated results using a model organism such as a mouse, and introducing a knock out or activating mutation into the candidate gene. The effect of the mutation can then be ascertained by seeing if cancer develops and measuring the invasiveness of the cancer.

A xenograft, in which human cells are put into a mouse without an immune system so that the human cells are not rejected, can also be used to measure how the targeted mutated gene affects the phenotype of the model organism.

8.2.2 Sub-region analysis

Evolutionary analysis in this project has demonstrated the power to be able to detect positive selection on whole genes. However this analysis has not revealed which regions of significantly mutated genes specifically are driving cancer.

To refine this gene-based analysis to uncover the regional patterns of SNV selection within genes, positive selection can be measured at the sub-region level to investigate

where exactly driver mutations are preferentially found in driver genes. This can be done by splitting the gene into functional sub-regions, for example using short linear motifs and globular protein domains as the units of analysis. Short linear motifs are generally situated in disordered regions of the genome and can be split into two high level classes: modification sites and ligand binding sites. The majority of known interaction partners of short linear motifs are globular protein domains.

By looking more closely at where driver mutations occur within cancer genes, the function of the gene can be more specifically related to its molecular consequence in cancer.

8.2.2.1 Motivation

The importance of partitioning genes by functional sub-regions is motivated by evolutionary results from PAML, which have shown certain genes to exhibit a significant p-value supporting evidence of positive selection, but a contradictory omega value indicative of negative selection or neutrality. These genes are suspected to be undergoing both positive and negative selection (in different regions of the gene), which could be why they have previously been missed as cancer genes (significantly enriched with driver mutations) in earlier evolutionary-based screens for driver genes. It is hoped that this type of refined analysis will help to separate out the potentially competing signals of selection within these genes that are confounding whole-gene based analysis, by focusing on specific functional regions to find where specifically positive selection is acting in the gene.

8.2.2.2 Globular protein domains: protein kinase domains

The protein kinase domain is an example of a globular protein domain, and is the most commonly found domain in known cancer genes [Greenman et al., 2007]. The kinase domain has also shown to be an important therapeutic target in several classes of human cancer with inhibitors of mutated protein kinases showing positive results in cancer

treatment, for example the use of the kinase inhibitors trastuzumab to treat breast cancer and inatimib to treat chronic myelogenous leukemia [Garber, 2006]. Therefore this domain is a good subject for sub-domain analysis, and would be investigated first from the set of domain coordinates previously obtained from Ensembl. Preparation for domain-based analysis has already begun as described in Chapter 2 splitting MAPK1 into kinase domain and non-kinase domain, however the annotated alignments are yet to be run through PAML (Figure 8.1).

Once in PAML, parameters such as substitution rate and omega ratio are co-estimated across the whole gene as well as in the region of interest within the gene (in this case the kinase domain) and in the remainder of the gene. This approach gives more free parameters to provide a better fit to the model, and therefore more power to detect selection compared with gene-based analysis, which is a benefit when using such a small dataset such as a single gene and a complex codon model as is used in codeml. The results from PAML indicate how strong the selection signals are in the specific domain of interest relative to the remainder of the protein in the unstructured regions. The more sequences available for this gene the more power the analysis has, so using all cancer-types in the TCGA or Lawrence dataset as was done in Chapter 4 would be best suited for this type of analysis.

Other domains that could be considered for sub-domain analysis are as follows: oxygen dependent degradation domains (ODDs) versus the rest of gene in HIF1; surface domains versus buried domains; or extracellular versus intracellular versus transmembrane domains in genes encoding membrane protein.

8.2.2.3 Short linear motifs: phosphorylation sites

Phosphorylation sites are an example of a modification site, and are known to be the targets of kinase domains. However less is known about the role of these sites in cancer than is known about kinase domains, which makes them an interesting modification site to study in this sub-region analysis. Preparation for modification site-based analysis has

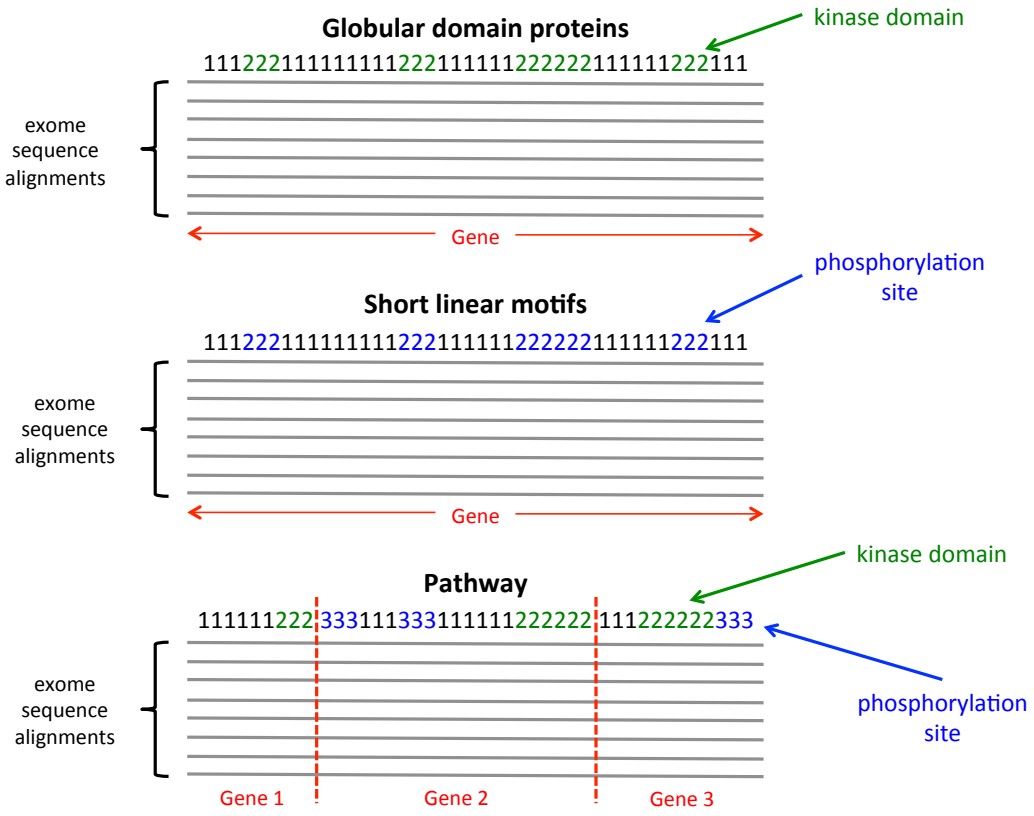


FIGURE 8.1: **Sub-region analysis in PAML.** PAML can be adapted to analyse alignments at the sub-region level by partitioning genes using functional region annotations. This figure shows a single gene annotated with kinase domain coordinates (indicated in green) obtained from Ensembl, and another instance where the gene is annotated with phosphorylation sites (indicated in blue) obtained from PhosPhoSite. The remainder of the gene is annotated with a 1 representing non-kinase domains or non-phosphorylation sites. PAML estimates an omega for each unique annotation. These regions can also be analysed in PAML over a whole pathway in the same way, by concatenating alignments across all genes in the pathway before estimating omega values for each annotated region.

also already begun as described in Chapter 2 partitioning TP53 by phosphorylation site, however again the annotated alignments are yet to be run through PAML (Figure 8.1).

This analysis could be extended by looking at de novo birth of phosphorylation sites (i.e. enrichment for mutations that look like they might create a phosphorylation site).

8.2.2.4 Pathway analysis

The sub-region approach described above can be adapted to detect positive selection in a specific region concatenated over multiple genes across a defined known pathway of interest. The aim is to uncover more about the mechanism occurring and understand exactly which part of a signalling cascade is affected by driver mutations in cancer. Pathway-based analysis increases the power at the sub-gene level as more data is used compared to looking at functional regions within a single gene, and detection power is therefore increased for small effects.

For example, phosphorylation sites can be concatenated across all genes in a pathway known to be involved in cancer, and separated from the non-phosphorylation sites of the pathway. PAML will then estimate an omega ratio for all phosphorylation sites and a separate one for the rest of the pathway, as well as an omega ratio over all genes in the whole pathway (Figure 8.1), to uncover which part of a known cancer implicated pathway is targeted by positive selection. This analysis can be refined by annotating several different regions simultaneously across all genes in a pathway, for example both kinase domains and phosphorylation sites within the JNK pathway [Greenman et al., 2007], and then asking PAML to calculate an omega for each region of interest. The results elucidate the specific origin of driver mutations in the pathway by detecting a stronger signal of selection either the kinase domain or phosphorylation site within the pathway. The MAPK/ERK signaling pathway is also a good candidate for the analysis of pathways involving kinase activity since it is known to have upstream mutations in cell-membrane-bound receptor tyrosine kinases such as EGFR, ERBB2, FGFR1, FGFR2,

FGFR3, PDGFRA and PDGFRB, as well as in downstream cytoplasmic components such as NF1, PTPN11, HRAS, KRAS, NRAS and BRAF [Stratton et al., 2009].

Pathway analysis could also just be applied to whole-genes, by grouping together genes involved in a particular pathway and concatenating all these gene alignments prior to evolutionary analysis to produce a single omega value for the pathway of interest.

KEGG [Kanehisa et al., 2014] is a resource that can be utilised to obtain knowledge of genes involved in particular pathways.

8.2.2.5 Novel approach

Neither the Lawrence et al. [2014] study nor the Kandoth et al. [2013] Pan-cancer study have addressed the hypothesis-driven question of where in genes and pathways is enriched for driver mutations detected in cancer, by dividing genes into functional sub-divisions. In both studies, and in many studies using the omega ratio to detect selection, genes have been used as the single unit of analysis. However, this has previously been addressed on a single gene in the work of Yang et al. [2003], where functional domains were partitioned in TP53 before using a codon model to measure selection using the omega ratio, akin to the methods used in this analysis. So as an innovation and to build on this work too, this approach could be applied at the whole genome-wide scale across all genes in the genome, using functional domains as well as modification sites concatenated across pathways as the units of analysis. Yang et al. [2003] also did not account for the varying coverage depth across the genome, so application of variant detection sensitivity correction in this context would be a further improvement to this novel approach making it more robust than the methods in Yang et al. [2003].

Additionally, data can be split using arbitrary data positions in known cancer genes to find novel candidate domains or sites involved in cancer.

8.2.3 Alternative evolutionary-based models

Other molecular adaptation models are available for evolutionary-based analysis such as the "sitewise likelihood-ratio" (SLR) [Massingham and Goldman, 2005] method which uses a site by site approach rather than a codon model (as is used in codeml), and can also be adapted to detect purifying selection. A comparison of the results from different maximum likelihood methods could be useful in seeing which has more power.

Results can be complemented with simulation studies and permutation analyses, which can act as a control set of results for evolutionary analyses.

8.2.4 Further ways to partition data

As well as tissue of origin, mutation spectra, functional sub-domains and pathways, data could also be partitioned on many other conditions to further understand the underlying biology of the cancer. For example, considering pre-existing mutations (germline), prognosis, age of patient, etc. and accounting for each of these factors prior to evolutionary analysis.

8.2.5 Improved mutation profiles

This analysis was intended to be a refinement of what has gone before, however more work is needed to further characterise mutation spectra in cancers and relate it to the path of cancer development.

In this analysis, only the six classes of single nucleotide mutations were considered. However, the classification system in Kandoth et al. [2013] could be adopted in order to improve these basic classifications before evolutionary analysis. For example, data could be stratified further based on the context of the mutation, for example considering CpG effects. Mutation rates are known to be elevated at CpG sites, and especially in cancer with certain cancers found to have a much higher mutation rate at CpG sites [Kandoth

et al., 2013]. Since threonine and serine are both phosphorylated and also contain CpG sites, investigating CpG effects would complement the sub-region analysis using phosphorylation sites. Using the phosphorylation data, we could assess the incidence of phosphorylation sites with CpG sites hit by driver mutations versus those without CpG sites. This style of analysis could indicate how the CpG specific mutation rate affects which genes are hit by driver mutations in cancer. For example, a CpG mutation in arginine produces a stop codon, and stop codons are known to occur frequently in the APC gene in colorectal cancer. We can ask if this high mutation rate of stop codons in APC in colorectal cancer is caused by the presence of a CpG dinucleotide. Mutation spectra (including CpG effects) across different cancer types has been addressed in Kandoth et al. [2013], however they have not addressed how the mutation spectrum influences which genes are hit by candidate driver mutations. The advantage is that Kandoth et al. [2013] analysis can be built on by using the same classifications but by then looking further using measures of selection to detect which genes are hit by mutation in different mutation spectra scenarios.

8.2.6 Meta-analysis

TCGA and Cancer Genome Consortium projects are continuously expanding and generating genomic data, so the addition of further cancer exomes across many different cancer types would increase the power of analysis. Whole-genome sequences are also available, and could be used to investigate the non-coding regions of the genome in cancer. Other major sequencing project datasets could also be incorporated, such as the Wellcome Trust Sanger Institute's Cancer Genome Project [Futreal et al., 2004]. This resource uses the human genome sequence and high-throughput mutation detection techniques to identify somatically acquired sequence variants and hence identify genes critical to the development of human cancers. One of their projects consists of the full coding sequence of 518 protein kinase genes over 210 diverse human tumour samples together with their matched somatic sequences, for both primary and cell-line samples [Greenman et al., 2007]. This dataset would be a good choice for the study of

sub-region analysis in kinase domains and phosphorylation sites, especially since each sample has a coding sequence of ∼1.3Mb of DNA which covers a large proportion of the genome.

8.2.7 Purifying selection

A more neglected category of genes in cancer studies are those whose normal function is required for cancer progression. These genes would be expected to exhibit strong purifying selection in evolutionary studies, due to mutations deleterious to the cancer being removed from these genes in the cancer population, showing evidence that these genes are conserved in cancer. For example, MTH1 is a gene required for the survival of many cancers, but it is non-essential in normal cells [Gad et al., 2014]. This would be a good candidate gene for a prototype of purifying selection in cancer genomes.

However, this is a high risk approach at the gene level due to background noise confounding selection signals. This has been observed in our results where genes are seen to have a significant p-value supporting positive selection, yet also an omega indicative of neutrality or negative selection. It is speculated that these genes are undergoing different modes of selection in different regions of the gene. Massingham and Goldman [2005] used evolutionary approaches to find the location of both positive and purifying selection acting in genes. Therefore, the sub-region analysis described above could be combined with a screen for purifying selection in order to identify conserved regions that may be important for cancer development.

8.2.8 INDELs and other types of mutation

INDELs have been explored briefly in this project in the FARP genes and in GBM, however they have not been investigated in an evolutionary context. This is because the models for this type of analysis do not exist in PAML. Lawrence et al. [2014] did however include INDELs in their study, since their models were able to deal with this type of mutation. INDELs are also more difficult to align and detect than SNVs, as

has been discovered in the FARP genes, so the pipeline currently in place would need to be adapted to deal with this type of mutation.

Large-scale structural deletions such as the deletions observed in FARP genes in this project would also be interesting to look at using evolutionary models. In this case whole-genome sequencing data would be required using physical coverage aided by the analysis of paired reads to first detect unexpected read pairing indicative of the presence of structural anomalies [Meyerson et al., 2010].

Investigating different types of mutation in the same cancers would give a more comprehensive understanding of the pattern of mutation profiles (not just limited to point mutations) and how these link to the cancer phenotype, and would provide more information to uncover the complex mechanisms underlying cancer progression.

8.3 Summary

To summarise, I have developed a protocol that is capable of rediscovering known cancer genes such as TP53, as proof of concept. I have also identified novel genes using this method that have not previously been identified as cancer driver genes. An interesting cancer-specific mutational profile has been identified in FARP1 and FARP2, which is not only a novel finding in cancer exomes but also highlights the difficulty in calling INDELs using current variant calling tools.

Appendix A

Variations between TCGA and Lawrence cancer type classifications

TABLE A.1: **Variations between TCGA and Lawrence cancer type classifications.** The cancer type name has been stated in this table where the definitions vary between TCGA and Lawrence classifications.

Cancer type abbreviation	Full cancer type name	
	TCGA	Lawrence
BLCA	Bladder urothelial carcinoma	Bladder
BRCA	Breast invasive carcinoma	Breast
CRC	Colorectal carcinoma	Colorectal
HNSC	Head and neck squamous cell carcinoma	Head and neck
KIRC	Kidney renal clear cell carcinoma	Kidney clear cell
OV	Ovarian serous cystadenocarcinoma	Ovarian
UCEC	Uterine corpus endometrial carcinoma	Endometrial

Appendix B

Example PAML control file for TP53

```
seqfile = ENSG00000141510.aln    * sequence data file name
treefile = ENSG00000141510.dnd     * tree structure file name
outfile = ENSG00000141510.codemlOUT * main result file name

noisy = 0 * 0,1,2,3,9: how much rubbish on the screen
  verbose = 0 * 0: concise output; 1: detailed output
runmode = 0 * 0: user tree;  1: semi-automatic;  2: automatic
      * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

seqtype = 1    * 1:codons; 2:AAs; 3:codons-->AAs
  CodonFreq = 2 * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table
ndata = 10
clock = 0      * 0:no clock, 1:clock; 2:local clock

aaDist = 0           * 0:equal, +:geometric; -:linear,
               * 1-6:G1974,Miyata,c,p,v,a; 7:AAClasses
aaRatefile = dat/jones.dat  * only used for aa seqs with model=empirical
 (_F)
                       * dayhoff.dat, jones.dat, wag.dat, mtmam.dat, or
  your own

```

```
model = 0 * models for codons:
        * 0:one, 1:b, 2:2 or more dN/dS ratios for branches
      * models for AAs or codon-translated AAs:
        * 0:poisson, 1:proportional, 2:Empirical
        * 3:Empirical+F, 6:FromCodon, 8:REVaa_0
                  * 9:REVaa(nr=189)

NSsites = 0 1 2 * 0:one w;1:neutral;2:selection; 3:discrete;4:freqs;
            * 5:gamma;6:2gamma;7:beta;8:beta&w;9:beta&gamma;
            * 10:beta&gamma+1; 11:beta&normal>1; 12:0&2normal>1;
                  * 13:3normal>0

icode = 0 * 0:universal code; 1:mammalian mt; 2-10:see below
Mgene = 0 * 0:rates, 1:separate

  fix_kappa = 0 * 1: kappa fixed, 0: kappa to be estimated
kappa = 2     * initial or fixed kappa
  fix_omega = 0   * 1: omega or omega_1 fixed, 0: estimate
omega = .4    * initial or fixed omega, for codons or codon-based AAs

fix_alpha = 1 * 0: estimate gamma shape parameter; 1: fix it at alpha
alpha = 0.    * initial or fixed alpha, 0:infinity (constant rate)
Malpha = 0    * different alphas for genes
ncatG = 8     * # of categories in dG of NSsites models

getSE = 0       * 0: don't want them, 1: want S.E.s of estimates
RateAncestor = 1    * (0,1,2): rates (alpha>0) or ancestral states (1 or
   2)

  Small_Diff = .5e-6
  cleandata = 0   * remove sites with ambiguity data (1:yes, 0:no)?
```

Listing B.1: Input control file used in codeml analysis in PAML for TP53 gene

Appendix C

FARP analysis: Perl code

```
#!/usr/bin/perl

=head1 NAME

removeDuplicateINDELs.pl

=head1 AUTHOR

Joanna Pethick (Joanna.Pethick@igmm.ed.ac.uk)

=head1 DESCRIPTION

Remove multiple transcripts for each INDEL in a gene in a patient, using
    the transcript with the most severe consequence only for each INDEL.

=cut

use strict;
use IO::File;
use Getopt::Long;

my $usage = qq{USAGE:
$0 [--help]
```

```
    --indels     FARP_indels
    --output     FARP_indels_unique
};

my $help = 0;
my $indels;
my $output;

GetOptions(
        'help'              => \$help,
        'indels=s'          => \$indels,
        'output=s'          => \$output,
) or die $usage;

if ($help || !$indels || !$output)
{
        print $usage;
        exit(0);
}

my $INDELS = new IO::File;
$INDELS->open($indels,"r") or die "Could not open $indels\n$!";

my $OUT = new IO::File;
$OUT->open($output,"w") or die "Could not open $output\n$!";

my %data;
my $current_line;
my $severity_max = -1;

#using ranked severity of consequences from snpEff
my %severity = ("UTR_3_PRIME" => 0,
      "UTR_5_PRIME" => 1,
            "UTR_3_DELETED" => 2,
            "UTR_5_DELETED" => 3,
      "CODON_CHANGE_PLUS_CODON_DELETION" => 4,
      "CODON_DELETION" => 5,
```

```
60        "CODON_CHANGE_PLUS_CODON_INSERTION" => 6,
61        "CODON_INSERTION" => 7,
62        "FRAME_SHIFT" => 8,
63              "EXON_DELETED" => 9,
64        "SPLICE_SITE_DONOR" => 10,
65        "SPLICE_SITE_ACCCEPTOR" => 11);
66
67 while (my $line = <$INDELS>)
68 {
69   chomp $line;
70   if ($line =~ /^gene/)
71     {
72         next;
73     }
74   my ($gene_name, $patient_id, $type, $chr, $pos, $ref, $alt, $cons,
     $disease, $cancer_genotype, $control_genotype, $cancer_genotype_qual,
     $control_genotype_qual, $gene, $transcript) = split(/\t/, $line);
75   my $variant = "$patient_id:$gene_name:$chr:$pos:$ref:$alt";
76   if ($current_line eq "$patient_id:$gene_name:$chr:$pos:$ref:$alt")
77   {
78     if ($severity{$cons} > $severity_max)
79     {
80       $severity_max = $severity{$cons};
81       $data{$variant} = $line;
82     }
83   }
84   else
85   {
86       $data{$variant} = $line;
87   }
88   $current_line = "$patient_id:$gene_name:$chr:$pos:$ref:$alt";
89   $severity_max = -1;
90 }
91 print $OUT "gene_name\tpatient_id\ttype\tchr\tpos\tref\talt\tcons\tdisease
     \tcancer_genotype\tcontrol_genotype\tcancer_genotype_qual\
     tcontrol_genotype_qual\tgene\ttranscript\n";
92 foreach my $variant (keys %data)
```

```
93 {
94   print $OUT "$data{$variant}\n";
95 }
```

Listing C.1: Perl code to find most severe consequence for each heterozygous cancer-specific INDEL mutation in FARP1 and FARP2 genes

Appendix D

FARP analysis: R code

```
longInsertions_1<-read.table("insertion_matches_farp1.unique", header = T)
longInsertions_1<-longInsertions_1[order(longInsertions_1$New_position_of_sequence),]
x11(width=16,height=7)
#use GRCh37 coordinates for where gene maps to in GRCh37 assembly
plot (longInsertions_1$Original_position_of_insertion, c(1:length(longInsertions_1$Original_position_of_insertion)), type="n", ann=FALSE, yaxt='n', xlim=c(98794816,99107430))
library("RMySQL")
dbh <- dbConnect(dbDriver("MySQL"), dbname="hg19", host="blackadder", user="myselect", password="singleton", client.flag=CLIENT_MULTI_STATEMENTS)
#use all transcript exons
qu<-'select name2, name, chrom, strand, txStart, txEnd, exonStarts, exonEnds from ensGene where name2="ENSG00000152767";'
qurt<-dbGetQuery(dbh,qu)
ymin <- 0
ymax <- 28
yv<-c(1:length(longInsertions_1$Original_position_of_insertion))
for (i in 1:length(qurt[,1]))
{
        start <- as.numeric(unlist(strsplit(qurt$exonStarts[i], ",")))
        end <- as.numeric(unlist(strsplit(qurt$exonEnds[i], ",")))
```

```
18
19          for (j in 1:length(start))
20          {
21                  y <- c(ymin, ymax, ymax, ymin)
22                  x <- c(start[j], start[j], end[j], end[j])
23                  polygon(x,y,col="lightblue", border="lightblue")
24          }
25  }
26
27  segments(longInsertions_1$Original_position_of_insertion, yv,
        longInsertions_1$New_position_of_sequence, yv, col="grey")
28
29  points(longInsertions_1$Original_position_of_insertion, c(1:length(
        longInsertions_1$Original_position_of_insertion)),col = "blue", pch =
        20, cex = 0.5)
30  points(longInsertions_1$New_position_of_sequence, c(1:length(
        longInsertions_1$Original_position_of_insertion)), col = "red", pch =
        20, cex = 0.5)
31  title(main="Long insertions (>=8nt) in FARP1 mapped back to hg19 reference
        genome")
32  title(xlab="Genomic position on chromosome 13 (GRCh37 assembly coordinates
        )")
33  legend("topleft",inset=0.05,c("Location of called insertion", "Location of
        insertion in reference"),col=c("red","blue"),pch=20,cex=0.8)
34  dev.copy2pdf(file="FARP1_map.pdf")
```

LISTING D.1: R code used to map long insertion (≥8nc) sequences in FARP1 back to the hg19 reference genome (script was modified for use on FARP2)

Appendix E

Unique TCGA SNVs filtered by most severe consequence: Perl code

```
#!/usr/bin/perl

=head1 NAME

removeMultipleTrans.pl

=head1 AUTHOR

Joanna Pethick (Joanna.Pethick@igmm.ed.ac.uk)

=head1 DESCRIPTION

Remove multiple transcripts for each mutation in a gene in a patient, by
    choosing the most severe consequence (non-synonymous). If a mutation
    appears in 2 overlapping genes, it is counted only once.

=cut

```

```
17 use strict;
18 use IO::File;
19 use Getopt::Long;
20
21 my $usage = qq{USAGE:
22 $0 [--help]
23     --tcga     all_patients or tcga_snvs_allCons
24     --output   all_patients_filtered or tcga_snvs_allCons_filtered
25 };
26
27 my $help = 0;
28 my $tcga;
29 my $output;
30
31 GetOptions(
32         'help'              => \$help,
33         'tcga=s'            => \$tcga,
34         'output=s'          => \$output,
35 ) or die $usage;
36
37 if ($help || !$tcga || !$output)
38 {
39         print $usage;
40         exit(0);
41 }
42
43 my $TCGA = new IO::File;
44 $TCGA->open($tcga,"r") or die "Could not open $tcga\n$!";
45 my $OUT = new IO::File;
46 $OUT->open($output,"w") or die "Could not open $output\n$!";
47 print $OUT "patient_id\tdisease\tdbsnp\t1kg_maf\t1kg_minor_allele\tchr\
    tpos\tref_allele\talt_allele\tconsequence\tgene\ttranscript\tpep_pos\n
    ";
48 my %data;
49 my $current_line;
50 my $most_severe;
51 my %severity = ("UTR_3_PRIME" => 0,
```

```
                    "UTR_5_PRIME" => 1,
                    "SYNONYMOUS_STOP" => 2,
                    "SYNONYMOUS_CODING" => 3,
                    "NON_SYNONYMOUS_START" => 4,
                    "SYNONYMOUS_START" => 5,
                    "NON_SYNONYMOUS_CODING" => 6,
                    "STOP_LOST" => 7,
                    "STOP_GAINED" => 8,
                    "START_LOST" => 9,
                    "SPLICE_SITE_DONOR" => 10,
                    "SPLICE_SITE_ACCEPTOR" => 11
                   );
while (my $line = <$TCGA>)
{
  chomp $line;
  if ($line =~ /^patient/)
    {
        next;
    }
  my ($patient, $disease, $dbsnp, $maf, $minor_allele, $chr, $position,
    $ref_allele, $alt_allele, $consequence, $gene, $transcript, $pep_pos)
    = split(/\t/, $line);
  my $variant = "$patient:$chr:$position:$ref_allele:$alt_allele";
  if ($current_line eq "$patient:$chr:$position:$ref_allele:$alt_allele")
  {
    if ($severity{$consequence} == 1)
    {
      $data{$variant} = $line;
    }
  }
  else
  {
      $data{$variant} = $line;
  }
  $current_line = "$patient:$chr:$position:$ref_allele:$alt_allele";
}
foreach my $variant (keys %data)
```

```
87 {
88   print $OUT "$data{$variant}\n";
89 }
```

LISTING E.1: Perl code used to filter the heterozygous cancer-specific TCGA SNVs to output the mutation on the transcript with the most severe consequence for each gene for each patient to obtain a set of unique SSNVs and for mutations occurring in overlapping genes the mutation has been counted once for patient/disease/genome-based analysis (script was modified for gene-based analysis to allow mutations in overlapping genes to be counted separately in each gene).

Bibliography

Alexandrov, L. B., S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, M. Imielinsk, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, and M. R. Stratton
2013. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421.

Alioto, T. S., S. Derdak, T. A. Beck, P. C. Boutros, L. Bower, I. Buchhalter, M. D. Eldridge, N. J. Harding, L. E. Heisler, E. Hovig, D. T. W. Jones, A. G. Lynch, S. Nakken, P. Ribeca, A.-S. Sertier, J. T. Simpson, P. Spellmen, P. Tarpey, L. Tonon, D. Vodk, T. N. Yamaguchi, S. B. Agullo, M. Dabad, R. E. Denroche, P. Ginsbach, S. C. Heath, E. Raineri, C. L. Anderson, B. Brors, R. Drews, R. Eils, A. Fujimoto, F. C. Giner, M. He, P. Hennings-Yeomans, B. Hutter, N. Jger, R. Kabbe, C. Kandoth,

S. Lee, L. Ltourneau, S. Ma, H. Nakagawa, N. Paramasivam, A.-M. Patch, M. Peto, M. Schlesner, S. Seth, D. Torrents, D. A. Wheeler, L. Xi, J. Zhang, D. S. Gerhard, V. Quesada, R. Valds-Mas, M. Gut, T. J. Hudson, J. D. McPherson, X. S. Puente, and I. G. Gut

2014. A comprehensive assessment of somatic mutation calling in cancer genomes. *bioRxiv*, P. 012997.

AmericanCancerSociety

2015. *Global Cancer Facts & Figures 3rd Edition.* Atlanta: American Cancer Society.

Ascierto, P. A., J. M. Kirkwood, J.-J. Grob, E. Simeone, A. M. Grimaldi, M. Maio, G. Palmieri, A. Testori, F. M. Marincola, and N. Mozzillo

2012. The role of braf v600 mutation in melanoma. *Journal of translational medicine*, 10:85.

Atefi, M., B. Titz, E. Avramis, C. Ng, D. Wong, A. Lassen, M. Cerniglia, H. Escuin-Ordinas, D. Foulad, B. Comin-Anduix, T. G. Graeber, and A. Ribas

2015. Combination of pan-raf and mek inhibitors in nras mutant melanoma. *Molecular cancer*, 14(1):27.

Bálint E, E. and K. H. Vousden

2001. Activation and activities of the p53 tumour suppressor protein. *British journal of cancer*, 85(12):1813–1823.

Bao, R., L. Huang, J. Andrade, W. Tan, W. A. Kibbe, H. Jiang, and G. Feng

2014. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer informatics*, 13(Suppl 2):67.

Baylin, S. B.

2005. Dna methylation and gene silencing in cancer. *Nature clinical practice. Oncology*, 2 Suppl 1:S4–11.

Benjamini, Y. and Y. Hochberg

1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57:289300.

Beroukhim, R., G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. C. Lee, J. H. Huang, S. Alexander, J. Du, T. Kau, R. K. Thomas, K. Shah, H. Soto, S. Perner, J. Prensner, R. M. Debiasi, F. Demichelis, C. Hatton, M. A. Rubin, L. A. Garraway, S. F. Nelson, L. Liau, P. S. Mischel, T. F. Cloughesy, M. Meyerson, T. A. Golub, E. S. Lander, I. K. Mellinghoff, and W. R. Sellers

2007. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America*, 104(50):20007–20012.

Beroukhim, R., C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, K. T. Mc Henry, R. M. Pinchback, A. H. Ligon, Y.-J. Cho, L. Haery, H. Greulich, M. Reich, W. Winckler, M. S. Lawrence, B. A. Weir, K. E. Tanaka, D. Y. Chiang, A. J. Bass, A. Loo, C. Hoffman, J. Prensner, T. Liefeld, Q. Gao, D. Yecies, S. Signoretti, E. Maher, F. J. Kaye, H. Sasaki, J. E. Tepper, J. A. Fletcher, J. Tabernero, J. Baselga, M.-S. Tsao, F. Demichelis, M. A. Rubin, P. A. Janne, M. J. Daly, C. Nucera, R. L. Levine, B. L. Ebert, S. Gabriel, A. K. Rustgi, C. R. Antonescu, M. Ladanyi, A. Letai, L. A. Garraway, M. Loda, D. G. Beer, L. D. True, A. Okamoto, S. L. Pomeroy, S. Singer, T. R. Golub, E. S. Lander, G. Getz, W. R. Sellers, and M. Meyerson

2010. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905.

Bestor, T., A. Laudano, R. Mattaliano, and V. Ingram

1988. Cloning and sequencing of a cdna encoding dna methyltransferase of mouse cells: the carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *Journal of molecular biology*, 203(4):971–983.

Betteridge, D. J.

2000. What is oxidative stress? *Metabolism: clinical and experimental*, 49(2 Suppl 1):3–8.

Bignell, G., R. Smith, C. Hunter, P. Stephens, H. Davies, C. Greenman, J. Teague, A. Butler, S. Edkins, C. Stevens, S. O'Meara, A. Parker, T. Avis, S. Barthorpe,

L. Brackenbury, G. Buck, J. Clements, J. Cole, E. Dicks, K. Edwards, S. Forbes, M. Gorton, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, D. Jones, V. Kosmidou, R. Laman, R. Lugg, A. Menzies, J. Perry, R. Petty, K. Raine, R. Shepherd, A. Small, H. Solomon, Y. Stephens, C. Tofts, J. Varian, A. Webb, S. West, S. Widaa, A. Yates, A. J. M. Gillis, H. J. Stoop, R. J. H. L. M. van Gurp, J. W. Oosterhuis, L. H. J. Looijenga, P. A. Futreal, R. Wooster, and M. R. Stratton

2006. Sequence analysis of the protein kinase gene family in human testicular germ-cell tumors of adolescents and adults. *Genes, chromosomes & cancer*, 45(1):42–46.

Bignell, G. R., C. D. Greenman, H. Davies, A. P. Butler, S. Edkins, J. M. Andrews, G. Buck, L. Chen, D. Beare, C. Latimer, S. Widaa, J. Hinton, C. Fahey, B. Fu, S. Swamy, G. L. Dalgliesh, B. T. Teh, P. Deloukas, F. Yang, P. J. Campbell, P. A. Futreal, and M. R. Stratton

2010. Signatures of mutation and selection in the cancer genome. *Nature*, 463(7283):893–898.

Bignell, G. R., J. Huang, J. Greshock, S. Watt, A. Butler, S. West, M. Grigorova, K. W. Jones, W. Wei, M. R. Stratton, P. A. Futreal, B. Weber, M. H. Shapero, and R. Wooster

2004. High-resolution analysis of dna copy number using oligonucleotide microarrays. *Genome research*, 14(2):287–295.

Boulton, S. J.

2006. Cellular functions of the brca tumour-suppressor proteins. *Biochemical Society transactions*, 34(Pt 5):633–645.

Boulton, S. J.

2010. Dna repair: Decision at the break point. *Nature*, 465(7296):301–302.

Bouwman, P., A. Aly, J. M. Escandell, M. Pieterse, J. Bartkova, H. van der Gulden, S. Hiddingh, M. Thanasoula, A. Kulkarni, Q. Yang, B. G. Haffty, J. Tommiska, C. Blomqvist, R. Drapkin, D. J. Adams, H. Nevanlinna, J. Bartek, M. Tarsounas,

S. Ganesan, and J. Jonkers

2010. 53bp1 loss rescues brca1 deficiency and is associated with triple-negative and brca-mutated breast cancers. *Nature structural & molecular biology*, 17(6):688–695.

Bunting, S. F., E. Callén, N. Wong, H.-T. Chen, F. Polato, A. Gunn, A. Bothmer, N. Feldhahn, O. Fernandez-Capetillo, L. Cao, X. Xu, C.-X. Deng, T. Finkel, M. Nussenzweig, J. M. Stark, and A. Nussenzweig

2010. 53bp1 inhibits homologous recombination in brca1-deficient cells by blocking resection of dna breaks. *Cell*, 141(2):243–254.

Bunting, S. F. and A. Nussenzweig

2013. End-joining, translocations and cancer. *Nature reviews. Cancer*, 13(7):443–454.

Burrell, R. A., N. McGranahan, J. Bartek, and C. Swanton

2013. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345.

Cancer Genome Atlas Research Network

2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068.

Cancer Genome Atlas Research Network

2011. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615.

Cancer Research UK

2014. *Cancer Statistics Report: Cancer Survival in the UK up to 2011.*

Carter, H., S. Chen, L. Isik, S. Tyekucheva, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, and R. Karchin

2009. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer research*, 69(16):6660–6667.

Ceccaldi, R., J. C. Liu, R. Amunugama, I. Hajdu, B. Primack, M. I. R. Petalcorin, K. W. O'Connor, P. A. Konstantinopoulos, S. J. Elledge, S. J. Boulton, T. Yusufzai,

and A. D. D'Andrea

2015. Homologous-recombination-deficient tumours are dependent on polθ-mediated repair. *Nature*.

Chang, H., D. G. Jackson, P. S. Kayne, P. B. Ross-Macdonald, R.-P. Ryseck, and N. O. Siemers

2011. Exome sequencing reveals comprehensive genomic alterations across eight cancer cell lines. *PloS one*, 6(6):e21097.

Chilamakuri, C. S. R., S. Lorenz, M.-A. Madoui, D. Vodák, J. Sun, E. Hovig, O. Myklebost, and L. A. Meza-Zepeda

2014. Performance comparison of four exome capture systems for deep sequencing. *BMC genomics*, 15(1):449.

Chung, W., J. Bondaruk, J. Jelinek, Y. Lotan, S. Liang, B. Czerniak, and J.-P. J. Issa

2011. Detection of bladder cancer using novel dna methylation biomarkers in urine sediments. *Cancer Epidemiology Biomarkers & Prevention*, 20(7):1483–1491.

Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden

2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92.

Cleary, A. S., T. L. Leonard, S. A. Gestl, and E. J. Gunther

2014. Tumour cell heterogeneity maintained by cooperating subclones in wnt-driven mammary cancers. *Nature*, 508(7494):113–117.

Cleaver, J. E.

2005. Cancer in xeroderma pigmentosum and related disorders of dna repair. *Nature reviews. Cancer*, 5(7):564–573.

Consortium, . G. P. et al.

2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Consortium, I. H. . et al.

2010. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58.

Consortium, U. et al.

2014. Uniprot: a hub for protein information. *Nucleic acids research*, P. gku989.

Croce, C. M.

2008. Oncogenes and cancer. *The New England journal of medicine*, 358(5):502–511.

Damore, J. A. and J. Gore

2011. A slowly evolving host moves first in symbiotic interactions. *Evolution; international journal of organic evolution*, 65(8):2391–2398.

Darwin, C. and A. R. Wallace

1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. In *Proceedings of the Linnean Society of London*, volume 3, Pp. 45–62, London, UK. The Linnean Society.

de Boer, J. and J. H. Hoeijmakers

2000. Nucleotide excision repair and human syndromes. *Carcinogenesis*, 21(3):453–460.

de Cavanagh, E., A. E. Honegger, E. Hofer, R. H. Bordenave, E. O. Bullorsky, N. A. Chasseing, and C. Fraga

2002. Higher oxidation and lower antioxidant levels in peripheral blood plasma and bone marrow plasma from advanced cancer patients. *Cancer*, 94(12):3247–3251.

de Miranda, N. F., F. J. Hes, T. van Wezel, and H. Morreau

2012. Role of the microenvironment in the tumourigenesis of microsatellite unstable and mutyh-associated polyposis colorectal cancers. *Mutagenesis*, 27(2):247–253.

Deininger, M., E. Buchdunger, and B. J. Druker

2005. The development of imatinib as a therapeutic agent for chronic myeloid leukemia. *Blood*, 105(7):2640–2653.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly

2011. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491–498.

DeVita, V. T., T. S. Lawrence, and S. A. Rosenberg

2010. *Cancer: principles and practice of oncology-advances in oncology*, volume 1. Lippincott Williams & Wilkins.

Dong, J. T.

2001. Chromosomal deletions and tumor suppressor genes in prostate cancer. *Cancer metastasis reviews*, 20(3-4):173–193.

Dow, L. E., J. Fisher, K. P. O'Rourke, A. Muley, E. R. Kastenhuber, G. Livshits, D. F. Tschaharganeh, N. D. Socci, and S. W. Lowe

2015. Inducible in vivo genome editing with crispr-cas9. *Nature biotechnology*, 33(4):390–394.

Du, J., Y. Shi, Y. Pan, X. Jin, C. Liu, N. Liu, Q. Han, Y. Lu, T. Qiao, and D. Fan

2005. Regulation of multidrug resistance by ribosomal protein l6 in gastric cancer cells. *Cancer biology & therapy*, 4(2):242–247.

Dürst, M., L. Gissmann, H. Ikenberg, and H. zur Hausen

1983. A papillomavirus dna from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. *Proceedings of the National Academy of Sciences of the United States of America*, 80(12):3812–3815.

Dutt, A., H. B. Salvesen, T.-H. Chen, A. H. Ramos, R. C. Onofrio, C. Hatton, R. Nicoletti, W. Winckler, R. Grewal, M. Hanna, et al.

2008. Drug-sensitive fgfr2 mutations in endometrial carcinoma. *Proceedings of the National Academy of Sciences*, 105(25):8713–8717.

Eads, C. A., K. D. Danenberg, K. Kawakami, L. B. Saltz, P. V. Danenberg, and P. W. Laird

1999. Cpg island hypermethylation in human colorectal tumors is not associated with dna methyltransferase overexpression. *Cancer research*, 59(10):2302–2306.

Eden, E., R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini

2009. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC bioinformatics*, 10:48.

Felsenstein, J.

2005. Phylip: Phylogenetic inference program. *University of Washington:Seattle*, Version 3.6.

Fernald, G. H., E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman

2011. Bioinformatics challenges for personalized medicine. *Bioinformatics (Oxford, England)*, 27(13):1741–1748.

Fishel, R., M. K. Lescoe, M. R. Rao, N. G. Copeland, N. A. Jenkins, J. Garber, M. Kane, and R. Kolodner

1994. The human mutator gene homolog msh2 and its association with hereditary nonpolyposis colon cancer. *Cell*, 77(1):1 p following 166.

Flanagan, S. E., A.-M. Patch, and S. Ellard

2010. Using sift and polyphen to predict loss-of-function and gain-of-function mutations. *Genetic testing and molecular biomarkers*, 14(4):533–537.

Fletcher, M. N. C., M. A. A. Castro, X. Wang, I. de Santiago, M. O'Reilly, S.-F. Chin, O. M. Rueda, C. Caldas, B. A. J. Ponder, F. Markowetz, and K. B. Meyer

2013. Master regulators of fgfr2 signalling and breast cancer risk. *Nature communications*, 4:2464.

Flicek, P., I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. García-Girón, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A. K. Kähäri, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W. M. McLaren, M. Muffato, R. Nag,

B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T. J. P. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa, and S. M. J. Searle

2013. Ensembl 2013. *Nucleic acids research*, 41(Database issue):D48–D55.

Flicek, P., M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. Hunt, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, E. Kulesha, F. J. Martin, T. Maurel, W. M. McLaren, D. N. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, M. Ruffier, D. Sheppard, K. Taylor, A. Thormann, S. J. Trevanion, A. Vullo, S. P. Wilder, M. Wilson, A. Zadissa, B. L. Aken, E. Birney, F. Cunningham, J. Harrow, J. Herrero, T. J. P. Hubbard, R. Kinsella, M. Muffato, A. Parker, G. Spudich, A. Yates, D. R. Zerbino, and S. M. J. Searle

2014. Ensembl 2014. *Nucleic acids research*, 42(Database issue):D749–D755.

Fodde, R.

2002. The apc gene in colorectal cancer. *European journal of cancer (Oxford, England : 1990)*, 38(7):867–871.

Forbes, S. A., N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton, and P. A. Futreal

2011. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic acids research*, 39(Database issue):D945–D950.

Forbes, S. A., G. Tang, N. Bindal, S. Bamford, E. Dawson, C. Cole, C. Y. Kok, M. Jia, R. Ewing, A. Menzies, J. W. Teague, M. R. Stratton, and P. A. Futreal

2010. Cosmic (the catalogue of somatic mutations in cancer): a resource to investigate acquired mutations in human cancer. *Nucleic acids research*, 38(Database issue):D652–D657.

Forment, J. V., A. Kaidi, and S. P. Jackson

2012. Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nature reviews. Cancer*, 12(10):663–670.

Futreal, P. A., L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton

2004. A census of human cancer genes. *Nature reviews. Cancer*, 4(3):177–183.

Gad, H., T. Koolmeister, A.-S. Jemth, S. Eshtad, S. A. Jacques, C. E. Ström, L. M. Svensson, N. Schultz, T. Lundbäck, B. O. Einarsdottir, A. Saleh, C. Göktürk, P. Baranczewski, R. Svensson, R. P.-A. Berntsson, R. Gustafsson, K. Strömberg, K. Sanjiv, M.-C. Jacques-Cordonnier, M. Desroses, A.-L. Gustavsson, R. Olofsson, F. Johansson, E. J. Homan, O. Loseva, L. Bräutigam, L. Johansson, A. Höglund, A. Hagenkort, T. Pham, M. Altun, F. Z. Gaugaz, S. Vikingsson, B. Evers, M. Henriksson, K. S. A. Vallin, O. A. Wallner, L. G. J. Hammarström, E. Wiita, I. Almlöf, C. Kalderén, H. Axelsson, T. Djureinovic, J. C. Puigvert, M. Häggblad, F. Jeppsson, U. Martens, C. Lundin, B. Lundgren, I. Granelli, A. J. Jensen, P. Artursson, J. A. Nilsson, P. Stenmark, M. Scobie, U. W. Berglund, and T. Helleday

2014. Mth1 inhibition eradicates cancer by preventing sanitation of the dntp pool. *Nature*, 508(7495):215–221.

Gambacorti-Passerini, C.

2008. Part i: Milestones in personalised medicine–imatinib. *The Lancet. Oncology*, 9(6):600.

Garber, K.

2002. Synthetic lethality: killing cancer with cancer. *Journal of the National Cancer Institute*, 94(22):1666–1668.

Garber, K.

2006. The second wave in kinase cancer drugs. *Nature biotechnology*, 24(2):127–130.

Gibbs, R. A., J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen, et al.

2003. The international hapmap project. *Nature*, 426(6968):789–796.

Goel, A. and C. R. Boland

2012. Epigenetics of colorectal cancer. *Gastroenterology*, 143(6):1442–1460.

Goldman, N. and Z. Yang

1994. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular biology and evolution*, 11(5):725–736.

Gonzalez-Perez, A., A. Jene-Sanz, and N. Lopez-Bigas

2013. The mutational landscape of chromatin regulatory factors across 4,623 tumor samples. *Genome biology*, 14(9):r106.

Greenman, C., P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y.-E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M.-H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton

2007. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158.

Greenman, C., R. Wooster, P. A. Futreal, M. R. Stratton, and D. F. Easton

2006. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, 173(4):2187–2198.

Hanahan, D. and R. A. Weinberg

2000. The hallmarks of cancer. *cell*, 100(1):57–70.

Hanahan, D. and R. A. Weinberg

2011. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674.

He, X., Y.-C. Kuo, T. J. Rosche, and X. Zhang

2013. Structural basis for autoinhibition of the guanine nucleotide exchange factor farp2. *Structure*, 21(3):355–364.

Heitzer, E. and I. Tomlinson

2014. Replicative dna polymerase mutations in cancer. *Current opinion in genetics & development*, 24:107–113.

Hollander, M. C., G. M. Blumenthal, and P. A. Dennis

2011. Pten loss in the continuum of common cancers, rare syndromes and mouse models. *Nature Reviews Cancer*, 11(4):289–301.

Hollstein, M., K. Rice, M. S. Greenblatt, T. Soussi, R. Fuchs, T. Sørlie, E. Hovig, B. Smith-Sørensen, R. Montesano, and C. C. Harris

1994. Database of p53 gene somatic mutations in human tumors and cell lines. *Nucleic acids research*, 22(17):3551–3555.

Hornbeck, P. V., J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham, and M. Sullivan

2012. Phosphositeplus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic acids research*, 40(Database issue):D261–D270.

Hornef, N., H. Olbrich, J. Horvath, M. A. Zariwala, M. Fliegauf, N. T. Loges, J. Wildhaber, P. G. Noone, M. Kennedy, S. E. Antonarakis, J.-L. Blouin, L. Bartoloni, T. Nüsslein, P. Ahrens, M. Griese, H. Kuhl, R. Sudbrak, M. R. Knowles, R. Reinhardt, and H. Omran

2006. Dnah5 mutations are a common cause of primary ciliary dyskinesia with outer dynein arm defects. *American journal of respiratory and critical care medicine*, 174(2):120–126.

Hu, J. and P. C. Ng

2013. Sift indel: predictions for the functional effects of amino acid insertions/deletions in proteins. *PloS one*, 8(10):e77940.

Jain, V. K. and N. C. Turner

2012. Challenges and opportunities in the targeting of fibroblast growth factor receptors in breast cancer. *Breast cancer research : BCR*, 14(3):208.

Jukes, T. H. and C. R. Cantor

1969. Evolution of protein molecules. *Mammalian protein metabolism*, 3:21–132.

Kaelin, W. G.

2004. The von hippel-lindau tumor suppressor gene and kidney cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 10(18 Pt 2):6290S–6295S.

Kaelin, W. G.

2005. The concept of synthetic lethality in the context of anticancer therapy. *Nature reviews cancer*, 5(9):689–698.

Kaelin, W. G.

2010. New cancer targets emerging from studies of the von hippel-lindau tumor suppressor protein. *Annals of the New York Academy of Sciences*, 1210:1–7.

Kandoth, C., M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, M. D. M. Leiserson, C. A. Miller, J. S. Welch, M. J. Walter, M. C. Wendl, T. J. Ley, R. K. Wilson, B. J. Raphael, and L. Ding

2013. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339.

Kanehisa, M., S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe

2014. Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic acids research*, 42(Database issue):D199–D205.

Kasprzyk, A.

2011. Biomart: driving a paradigm change in biological data management. *Database : the journal of biological databases and curation*, 2011:bar049.

Koss, L. G.

2007. The mystery of chromosomal translocations in cancer. *Cytogenetic and genome research*, 118(2-4):247–251.

Kumar, R., J. Manning, H. E. Spendlove, G. Kremmidiotis, R. McKirdy, J. Lee, D. N. Millband, K. M. Cheney, M. R. Stampfer, P. P. Dwivedi, et al.

2006. Znf652, a novel zinc finger protein, interacts with the putative breast tumor suppressor cbfa2t3 to repress transcription. *Molecular cancer research*, 4(9):655–665.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al.

2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lange, S. S., K.-i. Takata, and R. D. Wood

2011. Dna polymerases and cancer. *Nature reviews. Cancer*, 11(2):96–110.

Lascorz, J., A. Försti, B. Chen, S. Buch, V. Steinke, N. Rahner, E. Holinski-Feder, M. Morak, H. K. Schackert, H. Görgens, et al.

2010. Genome-wide association study for colorectal cancer identifies risk polymorphisms in german familial cases and implicates mapk signalling pathways in disease susceptibility. *Carcinogenesis*, 31(9):1612–1619.

Lawrence, M. S., P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander, and G. Getz

2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501.

Lawrence, M. S., P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman,

A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortés, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D.-A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, R. S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz

2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.

Lee, E., R. Iskow, L. Yang, O. Gokcumen, P. Haseley, L. J. Luquette, J. G. Lohr, C. C. Harris, L. Ding, R. K. Wilson, D. A. Wheeler, R. A. Gibbs, R. Kucherlapati, C. Lee, P. V. Kharchenko, P. J. Park, and Cancer Genome Atlas Research Network

2012. Landscape of somatic retrotransposition in human cancers. *Science (New York, N.Y.)*, 337(6097):967–971.

Leibeling, D., P. Laspe, and S. Emmert

2006. Nucleotide excision repair and cancer. *Journal of molecular histology*, 37(5-7):225–238.

Lengauer, C., K. W. Kinzler, and B. Vogelstein

1998. Genetic instabilities in human cancers. *Nature*, 396(6712):643–649.

Levine, A. J.

1997. p53, the cellular gatekeeper for growth and division. *Cell*, 88(3):323–331.

Ley, T. J., L. Ding, M. J. Walter, M. D. McLellan, T. Lamprecht, D. E. Larson, C. Kandoth, J. E. Payton, J. Baty, J. Welch, C. C. Harris, C. F. Lichti, R. R. Townsend, R. S. Fulton, D. J. Dooling, D. C. Koboldt, H. Schmidt, Q. Zhang, J. R. Osborne, L. Lin, M. O'Laughlin, J. F. McMichael, K. D. Delehaunty, S. D. McGrath, L. A. Fulton, V. J. Magrini, T. L. Vickery, J. Hundal, L. L. Cook, J. J. Conyers,

G. W. Swift, J. P. Reed, P. A. Alldredge, T. Wylie, J. Walker, J. Kalicki, M. A. Watson, S. Heath, W. D. Shannon, N. Varghese, R. Nagarajan, P. Westervelt, M. H. Tomasson, D. C. Link, T. A. Graubert, J. F. DiPersio, E. R. Mardis, and R. K. Wilson

2010. Dnmt3a mutations in acute myeloid leukemia. *The New England journal of medicine*, 363(25):2424–2433.

Li, H. and R. Durbin

2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup

2009. The sequence alignment/map format and samtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079.

Li, J., C. Yen, D. Liaw, K. Podsypanina, S. Bose, S. I. Wang, J. Puc, C. Miliaresis, L. Rodgers, R. McCombie, et al.

1997. Pten, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *science*, 275(5308):1943–1947.

Loeb, L. A.

2001. A mutator phenotype in cancer. *Cancer research*, 61(8):3230–3239.

Lunter, G. and M. Goodson

2011. Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome research*, 21(6):936–939.

Massingham, T. and N. Goldman

2005. Detecting amino acid sites under positive selection and purifying selection. *Genetics*, 169(3):1753–1762.

Masson, N. and P. J. Ratcliffe

2014. Hypoxia signaling pathways in cancer metabolism: the importance of co-selecting interconnected physiological pathways. *Cancer & metabolism*, 2(1):3.

Mateos-Gomez, P. A., F. Gong, N. Nair, K. M. Miller, E. Lazzerini-Denchi, and A. Sfeir
2015. Mammalian polymerase θ promotes alternative nhej and suppresses recombination. *Nature*.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo
2010. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303.

McLaren, W., B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham
2010. Deriving the consequences of genomic variants with the ensembl api and snp effect predictor. *Bioinformatics (Oxford, England)*, 26(16):2069–2070.

Meyerson, M., S. Gabriel, and G. Getz
2010. Advances in understanding cancer genomes through second-generation sequencing. *Nature reviews. Genetics*, 11(10):685–696.

Meynert, A. M., M. Ansari, D. R. FitzPatrick, and M. S. Taylor
2014. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC bioinformatics*, 15:247.

Meynert, A. M., L. S. Bicknell, M. E. Hurles, A. P. Jackson, and M. S. Taylor
2013. Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC bioinformatics*, 14:195.

Milinkovic, V., J. Bankovic, M. Rakic, N. Milosevic, T. Stankovic, M. Jokovic, Z. Milosevic, M. Skender-Gazibara, A. Podolski-Renic, M. Pesic, et al.
2012. Genomic instability and p53 alterations in patients with malignant glioma. *Experimental and molecular pathology*, 93(2):200–206.

Mitchell, A., H.-Y. Chang, L. Daugherty, M. Fraser, S. Hunter, R. Lopez, C. McAnulla, C. McMenamin, G. Nuka, S. Pesseat, et al.
2014. The interpro protein families database: the classification resource after 15 years. *Nucleic acids research*, P. gku1243.

Mitelman, F., B. Johansson, and F. Mertens

2007. The impact of translocations and gene fusions on cancer causation. *Nature reviews. Cancer*, 7(4):233–245.

Müller, A. and R. Fishel

2002. Mismatch repair and the hereditary non-polyposis colorectal cancer syndrome (hnpcc). *Cancer investigation*, 20(1):102–109.

Muller, P. A. and K. H. Vousden

2014. Mutant p53 in cancer: new functions and therapeutic opportunities. *Cancer cell*, 25(3):304–317.

Mullighan, C. G., S. Goorha, I. Radtke, C. B. Miller, E. Coustan-Smith, J. D. Dalton, K. Girtman, S. Mathew, J. Ma, S. B. Pounds, X. Su, C.-H. Pui, M. V. Relling, W. E. Evans, S. A. Shurtleff, and J. R. Downing

2007. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature*, 446(7137):758–764.

Murchison, E. P.

2008. Clonally transmissible cancers in dogs and tasmanian devils. *Oncogene*, 27 Suppl 2:S19–S30.

Nambiar, M. and S. C. Raghavan

2011. How does dna break during chromosomal translocations? *Nucleic acids research*, 39(14):5813–5825.

Negrini, S., V. G. Gorgoulis, and T. D. Halazonetis

2010. Genomic instability–an evolving hallmark of cancer. *Nature reviews. Molecular cell biology*, 11(3):220–228.

Neve, R. M., K. Chin, J. Fridlyand, J. Yeh, F. L. Baehner, T. Fevr, L. Clark, N. Bayani, J.-P. Coppe, F. Tong, T. Speed, P. T. Spellman, S. DeVries, A. Lapuk, N. J. Wang, W.-L. Kuo, J. L. Stilwell, D. Pinkel, D. G. Albertson, F. M. Waldman, F. McCormick,

R. B. Dickson, M. D. Johnson, M. Lippman, S. Ethier, A. Gazdar, and J. W. Gray
2006. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer cell*, 10(6):515–527.

Ng, S. B., E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson, and J. Shendure
2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276.

Nguyen, L. S., T. Schneider, M. Rio, S. Moutton, K. Siquier-Pernet, F. Verny, N. Boddaert, I. Desguerre, A. Munich, J. L. Rosa, et al.
2015. A nonsense variant in herc1 is associated with intellectual disability, megalencephaly, thick corpus callosum and cerebellar atrophy. *European Journal of Human Genetics*.

Nielsen, R. and Z. Yang
1998. Likelihood models for detecting positively selected amino acid sites and applications to the hiv-1 envelope gene. *Genetics*, 148(3):929–936.

Nowak, K. J. and K. E. Davies
2004. Duchenne muscular dystrophy and dystrophin: pathogenesis and opportunities for treatment. *EMBO reports*, 5(9):872–876.

Oberley, T. D.
2002. Oxidative damage and cancer. *The American journal of pathology*, 160(2):403–408.

Olivier, M., M. Hollstein, and P. Hainaut
2010. Tp53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology*, 2(1):a001008.

Pabinger, S., A. Dander, M. Fischer, R. Snajder, M. Sperk, M. Efremova, B. Krabichler, M. R. Speicher, J. Zschocke, and Z. Trajanoski

2014. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, 15(2):256–278.

Paek, A., C.-H. Lee, and H. J. You

2014. A role of zinc-finger protein 143 for cancer cell migration and invasion through zeb1 and e-cadherin in colon cancer cells. *Molecular carcinogenesis*, 53(S1):E161–E168.

Patanè, M., P. Porrati, E. Bottega, S. Morosini, G. Cantini, V. Girgenti, A. Rizzo, M. Eoli, B. Pollo, F. L. Sciacca, et al.

2013. Frequency of nfkbia deletions is low in glioblastomas and skewed in glioblastoma neurospheres. *Molecular cancer*, 12(1):1–12.

Paul, M. K. and A. K. Mukhopadhyay

2004. Tyrosine kinase–role and significance in cancer. *International journal of medical sciences*, 1(2):101.

Peltomäki, P.

2001. Deficient dna mismatch repair: a common etiologic factor for colon cancer. *Human molecular genetics*, 10(7):735–740.

Pleasance, E. D., R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M.-L. Lin, G. R. Ordóñez, G. R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R. J. Carter, L. Chen, A. J. Cox, S. Edkins, P. I. Kokko-Gonzales, N. A. Gormley, R. J. Grocock, C. D. Haudenschild, M. M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L. J. Mudie, Z. Ning, T. Royce, O. B. Schulz-Trieglaff, A. Spiridou, L. A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M. T. Ross, P. J. Campbell, D. R. Bentley, P. A. Futreal, and M. R. Stratton

2010a. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–196.

Pleasance, E. D., P. J. Stephens, S. O'Meara, D. J. McBride, A. Meynert, D. Jones, M.-L. Lin, D. Beare, K. W. Lau, C. Greenman, I. Varela, S. Nik-Zainal, H. R. Davies, G. R. Ordoñez, L. J. Mudie, C. Latimer, S. Edkins, L. Stebbings, L. Chen, M. Jia, C. Leroy, J. Marshall, A. Menzies, A. Butler, J. W. Teague, J. Mangion, Y. A. Sun, S. F. McLaughlin, H. E. Peckham, E. F. Tsung, G. L. Costa, C. C. Lee, J. D. Minna, A. Gazdar, E. Birney, M. D. Rhodes, K. J. McKernan, M. R. Stratton, P. A. Futreal, and P. J. Campbell

2010b. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463(7278):184–190.

Polakis, P.

1999. The oncogenic activation of beta-catenin. *Current opinion in genetics & development*, 9(1):15–21.

Polakis, P., M. Hart, and B. Rubinfeld

1999. Defects in the regulation of beta-catenin in colorectal cancer. *Advances in experimental medicine and biology*, 470:23–32.

Preston, B. D., T. M. Albertson, and A. J. Herr

2010. Dna replication fidelity and cancer. In *Seminars in cancer biology*, volume 20, Pp. 281–293. Elsevier.

Puente, X. S., M. Pinyol, V. Quesada, L. Conde, G. R. Ordóñez, N. Villamor, G. Escaramis, P. Jares, S. Beà, M. González-Díaz, L. Bassaganyas, T. Baumann, M. Juan, M. López-Guerra, D. Colomer, J. M. C. Tubío, C. López, A. Navarro, C. Tornador, M. Aymerich, M. Rozman, J. M. Hernández, D. A. Puente, J. M. P. Freije, G. Velasco, A. Gutiérrez-Fernández, D. Costa, A. Carrió, S. Guijarro, A. Enjuanes, L. Hernández, J. Yagüe, P. Nicolás, C. M. Romeo-Casabona, H. Himmelbauer, E. Castillo, J. C. Dohm, S. de Sanjosé, M. A. Piris, E. de Alava, J. San Miguel, R. Royo, J. L. Gelpí, D. Torrents, M. Orozco, D. G. Pisano, A. Valencia, R. Guigó, M. Bayés, S. Heath, M. Gut, P. Klatt, J. Marshall, K. Raine, L. A. Stebbings, P. A. Futreal, M. R. Stratton, P. J. Campbell, I. Gut, A. López-Guillermo, X. Estivill, E. Montserrat,

C. López-Otín, and E. Campo

2011. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*, 475(7354):101–105.

Quinlan, A. R. and I. M. Hall

2010. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–842.

Quintero, O. A., W. C. Unrath, S. M. Stevens, U. Manor, B. Kachar, and C. M. Yengo

2013. Myosin 3a kinase activity is regulated by phosphorylation of the kinase domain activation loop. *Journal of Biological Chemistry*, 288(52):37126–37137.

R Core Team

2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabbani, B., M. Tekin, and N. Mahdieh

2014. The promise of whole-exome sequencing in medical genetics. *Journal of human genetics*, 59(1):5–15.

Rajith, B., C. Chakraborty, et al.

2013. Predicting the impact of deleterious mutations in the protein kinase domain of fgfr2 in the context of function, structure, and pathogenesisa bioinformatics approach. *Applied biochemistry and biotechnology*, 170(8):1853–1870.

Rao, S. K., J. Edwards, A. D. Joshi, I.-M. Siu, and G. J. Riggins

2010. A survey of glioblastoma genomic amplifications and deletions. *Journal of neuro-oncology*, 96(2):169–179.

Reddy, E. P., R. K. Reynolds, E. Santos, and M. Barbacid

1982. A point mutation is responsible for the acquisition of transforming properties by the t24 human bladder carcinoma oncogene.

Reintjes, N., Y. Li, A. Becker, E. Rohmann, R. Schmutzler, and B. Wollnik

2013. Activating somatic fgfr2 mutations in breast cancer. *PloS one*, 8(3):e60264.

Robertson, K. D.

2001. Dna methylation, methyltransferases, and cancer. *Oncogene*, 20(24):3139–3155.

Romero, O. A., M. Torres-Diz, E. Pros, S. Savola, A. Gomez, S. Moran, C. Saez, R. Iwakawa, A. Villanueva, L. M. Montuenga, et al.

2014. Max inactivation in small cell lung cancer disrupts myc–swi/snf programs and is synthetic lethal with brg1. *Cancer discovery*, 4(3):292–303.

Rosenbloom, K. R., J. Armstrong, G. P. Barber, J. Casper, H. Clawson, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, G. Hickey, A. S. Hinrichs, R. Hubley, D. Karolchik, K. Learned, B. T. Lee, C. H. Li, K. H. Miga, N. Nguyen, B. Paten, B. J. Raney, A. F. A. Smit, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn, and W. J. Kent

2015. The ucsc genome browser database: 2015 update. *Nucleic acids research*, 43(Database issue):D670–D681.

Ross, K. A.

2014. Coherent somatic mutation in autoimmune disease. *PloS one*, 9(7):e101093.

Rubin, A. F. and P. Green

2009. Mutation patterns in cancer genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21766–21770.

Samuels, D. C., L. Han, J. Li, S. Quanghu, T. A. Clark, Y. Shyr, and Y. Guo

2013. Finding the lost treasures in exome sequencing data. *Trends in Genetics*, 29(10):593–599.

Schuster-Böckler, B. and B. Lehner

2012. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, 488(7412):504–507.

Sharief, F. S., P. J. Vojta, P. A. Ropp, and W. C. Copeland

1999. Cloning and chromosomal mapping of the human dna polymerase θ (polq), the eighth human dna polymerase. *Genomics*, 59(1):90–96.

Shendure, J. and H. Ji

2008. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145.

Shiffman, M. L. and Y. Benhamou

2015. Cure of hcv related liver disease. *Liver international : official journal of the International Association for the Study of the Liver*, 35 Suppl 1:71–77.

Siegelin, M. D. and A. C. Borczuk

2014. Epidermal growth factor receptor mutations in lung adenocarcinoma. *Laboratory Investigation*, 94(2):129–137.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting

2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews. Genetics*, 15(2):121–132.

Singhi, A. D., A. Cimino-Mathews, R. B. Jenkins, F. Lan, S. R. Fink, H. Nassar, R. Vang, J. H. Fetting, J. Hicks, S. Sukumar, et al.

2012. Myc gene amplification is often acquired in lethal distant breast cancer metastases of unamplified primary tumors. *Modern Pathology*, 25(3):378–387.

Slater, G. S. C. and E. Birney

2005. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, 6:31.

Slattery, M. L., A. Lundgreen, and R. K. Wolff

2012. Map kinase genes and colon and rectal cancer. *Carcinogenesis*, 33(12):2398–2408.

Smith, M. L. and Y. R. Seo

2002. p53 regulation of dna excision repair pathways. *Mutagenesis*, 17(2):149–156.

Solary, E., O. A. Bernard, A. Tefferi, F. Fuks, and W. Vainchenker

2014. The ten-eleven translocation-2 (tet2) gene in hematopoiesis and hematopoietic diseases. *Leukemia*, 28(3):485–496.

Song, J., Z. Du, M. Ravasz, B. Dong, Z. Wang, and R. M. Ewing

2015. A protein interaction between beta-catenin and dnmt1 regulates wnt signaling and dna methylation in colorectal cancer cells. *Molecular Cancer Research*, Pp. molcanres–0644.

Stephens, P., S. Edkins, H. Davies, C. Greenman, C. Cox, C. Hunter, G. Bignell, J. Teague, R. Smith, C. Stevens, et al.

2005. A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nature genetics*, 37(6):590–592.

Stephens, P. J., C. D. Greenman, B. Fu, F. Yang, G. R. Bignell, L. J. Mudie, E. D. Pleasance, K. W. Lau, D. Beare, L. A. Stebbings, S. McLaren, M.-L. Lin, D. J. McBride, I. Varela, S. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, M. A. Quail, J. Burton, H. Swerdlow, N. P. Carter, L. A. Morsberger, C. Iacobuzio-Donahue, G. A. Follows, A. R. Green, A. M. Flanagan, M. R. Stratton, P. A. Futreal, and P. J. Campbell

2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):27–40.

Stephens, P. J., D. J. McBride, M.-L. Lin, I. Varela, E. D. Pleasance, J. T. Simpson, L. A. Stebbings, C. Leroy, S. Edkins, L. J. Mudie, C. D. Greenman, M. Jia, C. Latimer, J. W. Teague, K. W. Lau, J. Burton, M. A. Quail, H. Swerdlow, C. Churcher, R. Natrajan, A. M. Sieuwerts, J. W. M. Martens, D. P. Silver, A. Langerød, H. E. G. Russnes, J. A. Foekens, J. S. Reis-Filho, L. van 't Veer, A. L. Richardson, A.-L. Børresen-Dale, P. J. Campbell, P. A. Futreal, and M. R. Stratton

2009. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 462(7276):1005–1010.

Stratton, M. R.

2011. Exploring the genomes of cancer cells: progress and promise. *Science (New York, N.Y.)*, 331(6024):1553–1558.

Stratton, M. R., P. J. Campbell, and P. A. Futreal
2009. The cancer genome. *Nature*, 458(7239):719–724.

Strausberg, R. L., S. F. Greenhut, L. H. Grouse, C. F. Schaefer, and K. H. Buetow
2001. In silico analysis of cancer through the cancer genome anatomy project. *Trends in cell biology*, 11(11):S66–S71.

Supek, F. and B. Lehner
2015. Differential dna mismatch repair underlies mutation rate variation across the human genome. *Nature*.

Supek, F., B. Miñana, J. Valcárcel, T. Gabaldón, and B. Lehner
2014. Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, 156(6):1324–1335.

Tabin, C. J., S. M. Bradley, C. I. Bargmann, R. A. Weinberg, A. G. Papageorge, E. M. Scolnick, R. Dhar, D. R. Lowy, and E. H. Chang
1982. Mechanism of activation of a human oncogene. *Nature*, 300(5888):143–149.

Talavera, D., M. S. Taylor, and J. M. Thornton
2010. The (non)malignancy of cancerous amino acidic substitutions. *Proteins*, 78(3):518–529.

Talbot, S. J. and D. H. Crawford
2004. Viruses and tumours–an update. *European journal of cancer (Oxford, England : 1990)*, 40(13):1998–2005.

Tian, X., D. Sun, Y. Zhang, S. Zhao, H. Xiong, and J. Fang
2008. Zinc finger protein 278, a potential oncogene in human colorectal cancer. *Acta biochimica et biophysica Sinica*, 40(4):289–296.

Timmermann, B., M. Kerick, C. Roehr, A. Fischer, M. Isau, S. T. Boerno, A. Wunderlich, C. Barmeyer, P. Seemann, J. Koenig, M. Lappe, A. W. Kuss, M. Garshasbi, L. Bertram, K. Trappe, M. Werber, B. G. Herrmann, K. Zatloukal, H. Lehrach, and

M. R. Schweiger

2010. Somatic mutation profiles of msi and mss colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. *PloS one*, 5(12):e15661.

Tryka, K. A., L. Hao, A. Sturcke, Y. Jin, Z. Y. Wang, L. Ziyabari, M. Lee, N. Popova, N. Sharopova, M. Kimura, and M. Feolo

2014. Ncbi's database of genotypes and phenotypes: dbgap. *Nucleic acids research*, 42(Database issue):D975–D979.

Turner, T.

2013. Plot protein: visualization of mutations. *J. Clinical Bioinformatics*, 3:14.

UniProt Consortium

2014. Activities at the universal protein resource (uniprot). *Nucleic acids research*, 42(Database issue):D191–D198.

Urnov, F. D., E. J. Rebar, M. C. Holmes, H. S. Zhang, and P. D. Gregory

2010. Genome editing with engineered zinc finger nucleases. *Nature reviews. Genetics*, 11(9):636–646.

Vandin, F., E. Upfal, and B. J. Raphael

2012. De novo discovery of mutated driver pathways in cancer. *Genome research*, 22(2):375–385.

Veal, C. D., P. J. Freeman, K. Jacobs, O. Lancaster, S. Jamain, M. Leboyer, D. Albanes, R. R. Vaghela, I. Gut, S. J. Chanock, and A. J. Brookes

2012. A mechanistic basis for amplification differences between samples and between genome regions. *BMC genomics*, 13:455.

Vogelstein, B. and K. W. Kinzler

2004. Cancer genes and the pathways they control. *Nature medicine*, 10(8):789–799.

Voldborg, B. R., L. Damstrup, M. Spang-Thomsen, and H. S. Poulsen
1997. Epidermal growth factor receptor (egfr) and egfr mutations, function and possible role in clinical trials. *Annals of Oncology*, 8(12):1197–1206.

Waddell, N., M. Pajic, A.-M. Patch, D. K. Chang, K. S. Kassahn, P. Bailey, A. L. Johns, D. Miller, K. Nones, K. Quek, M. C. J. Quinn, A. J. Robertson, M. Z. H. Fadlullah, T. J. C. Bruxner, A. N. Christ, I. Harliwong, S. Idrisoglu, S. Manning, C. Nourse, E. Nourbakhsh, S. Wani, P. J. Wilson, E. Markham, N. Cloonan, M. J. Anderson, J. L. Fink, O. Holmes, S. H. Kazakoff, C. Leonard, F. Newell, B. Poudel, S. Song, D. Taylor, N. Waddell, S. Wood, Q. Xu, J. Wu, M. Pinese, M. J. Cowley, H. C. Lee, M. D. Jones, A. M. Nagrial, J. Humphris, L. A. Chantrill, V. Chin, A. M. Steinmann, A. Mawson, E. S. Humphrey, E. K. Colvin, A. Chou, C. J. Scarlett, A. V. Pinho, M. Giry-Laterriere, I. Rooman, J. S. Samra, J. G. Kench, J. A. Pettitt, N. D. Merrett, C. Toon, K. Epari, N. Q. Nguyen, A. Barbour, N. Zeps, N. B. Jamieson, J. S. Graham, S. P. Niclou, R. Bjerkvig, R. Grützmann, D. Aust, R. H. Hruban, A. Maitra, C. A. Iacobuzio-Donahue, C. L. Wolfgang, R. A. Morgan, R. T. Lawlor, V. Corbo, C. Bassi, M. Falconi, G. Zamboni, G. Tortora, M. A. Tempero, Australian Pancreatic Cancer Genome Initiative, A. J. Gill, J. R. Eshleman, C. Pilarsky, A. Scarpa, E. A. Musgrove, J. V. Pearson, A. V. Biankin, and S. M. Grimmond
2015. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*, 518(7540):495–501.

Wang, Y., A. Marino-Enriquez, R. R. Bennett, M. Zhu, Y. Shen, G. Eilers, J.-C. Lee, J. Henze, B. S. Fletcher, Z. Gu, et al.
2014. Dystrophin is a tumor suppressor in human cancers with myogenic programs. *Nature genetics*, 46(6):601–606.

Watson, I. R., K. Takahashi, P. A. Futreal, and L. Chin
2013. Emerging patterns of somatic mutations in cancer. *Nature Reviews Genetics*, 14(10):703–718.

Wei, X., V. Walia, J. C. Lin, J. K. Teer, T. D. Prickett, J. Gartner, S. Davis, NISC Comparative Sequencing Program, K. Stemke-Hale, M. A. Davies, J. E. Gershenwald,

W. Robinson, S. Robinson, S. A. Rosenberg, and Y. Samuels

2011. Exome sequencing identifies grin2a as frequently mutated in melanoma. *Nature genetics*, 43(5):442–446.

Weinstein, J. N., E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network, et al.

2013. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120.

Weir, B. A., M. S. Woo, G. Getz, S. Perner, L. Ding, R. Beroukhim, W. M. Lin, M. A. Province, A. Kraja, L. A. Johnson, K. Shah, M. Sato, R. K. Thomas, J. A. Barletta, I. B. Borecki, S. Broderick, A. C. Chang, D. Y. Chiang, L. R. Chirieac, J. Cho, Y. Fujii, A. F. Gazdar, T. Giordano, H. Greulich, M. Hanna, B. E. Johnson, M. G. Kris, A. Lash, L. Lin, N. Lindeman, E. R. Mardis, J. D. McPherson, J. D. Minna, M. B. Morgan, M. Nadel, M. B. Orringer, J. R. Osborne, B. Ozenberger, A. H. Ramos, J. Robinson, J. A. Roth, V. Rusch, H. Sasaki, F. Shepherd, C. Sougnez, M. R. Spitz, M.-S. Tsao, D. Twomey, R. G. W. Verhaak, G. M. Weinstock, D. A. Wheeler, W. Winckler, A. Yoshizawa, S. Yu, M. F. Zakowski, Q. Zhang, D. G. Beer, I. I. Wistuba, M. A. Watson, L. A. Garraway, M. Ladanyi, W. D. Travis, W. Pao, M. A. Rubin, S. B. Gabriel, R. A. Gibbs, H. E. Varmus, R. K. Wilson, E. S. Lander, and M. Meyerson

2007. Characterizing the cancer genome in lung adenocarcinoma. *Nature*, 450(7171):893–898.

Welsh, J. S.

2011. Contagious cancer. *The oncologist*, 16(1):1–4.

Whelan, S. and N. Goldman

2004. Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics*, 167(4):2027–2043.

Wooster, R., G. Bignell, J. Lancaster, S. Swift, S. Seal, J. Mangion, N. Collins, S. Gregory, C. Gumbs, and G. Micklem

1995. Identification of the breast cancer susceptibility gene brca2. *Nature*, 378(6559):789–792.

Xiao, W., X. Qu, X. Li, Y. Sun, H. Zhao, S. Wang, and X. Zhou

2015. Identification of commonly dysregulated genes in colorectal cancer by integrating analysis of rna-seq data and qrt-pcr validation. *Cancer gene therapy*, 22(5):278–284.

Xie, W.-H., B. Zhang, L.-H. Wang, C.-Y. Liu, C. Chang, B. Lei, and Y.-W. Wu

2015. Biodistribution characteristics and spect imaging of (99m)tc-ret and (99m)tc-reg in human lung cancer xenografts. *Cancer biotherapy & radiopharmaceuticals*.

Xu, Y., L. Diao, Y. Chen, Y. Liu, C. Wang, T. Ouyang, J. Li, T. Wang, Z. Fan, T. Fan, et al.

2013. Promoter methylation of brca1 in triple-negative breast cancer predicts sensitivity to adjuvant chemotherapy. *Annals of oncology*, P. mdt011.

Yamazaki, J., R. Taby, A. Vasanthakumar, T. Macrae, K. R. Ostler, L. Shen, H. M. Kantarjian, M. R. Estecio, J. Jelinek, L. A. Godley, and J.-P. J. Issa

2012. Effects of tet2 mutations on dna methylation in chronic myelomonocytic leukemia. *Epigenetics : official journal of the DNA Methylation Society*, 7(2):201–207.

Yang, L., R. Rau, and M. A. Goodell

2015. Dnmt3a in haematological malignancies. *Nature reviews. Cancer*, 15(3):152–165.

Yang, Z.

2007. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591.

Yang, Z. and R. Nielsen

1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of molecular evolution*, 46(4):409–418.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen

2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–449.

Yang, Z., S. Ro, and B. Rannala

2003. Likelihood models of somatic mutation and codon substitution in cancer genes. *Genetics*, 165(2):695–705.

Yang, Z., W. S. W. Wong, and R. Nielsen

2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular biology and evolution*, 22(4):1107–1118.

Yokota, J.

2000. Tumor progression and metastasis. *Carcinogenesis*, 21(3):497–503.

Youn, A. and R. Simon

2011. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*, 27(2):175–181.

Yu, J., Q. Liang, J. Wang, Y. Cheng, S. Wang, T. Poon, M. Go, Q. Tao, Z. Chang, and J. Sung

2013. Zinc-finger protein 331, a novel putative tumor suppressor, suppresses growth and invasiveness of gastric cancer. *Oncogene*, 32(3):307–317.

Zhang, C.-Z., M. L. Leibowitz, and D. Pellman

2013. Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. *Genes & development*, 27(23):2513–2530.

Zhang, J., J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, et al.

2011. International cancer genome consortium data portala one-stop shop for cancer genomics data. *Database*, 2011:bar026.

Zhao, Q., E. F. Kirkness, O. L. Caballero, P. A. Galante, R. B. Parmigiani, L. Edsall, S. Kuan, Z. Ye, S. Levy, A. T. R. Vasconcelos, B. Ren, S. J. de Souza, A. A. Camargo, A. J. G. Simpson, and R. L. Strausberg

2010. Systematic detection of putative tumor suppressor genes through the combined use of exome and transcriptome sequencing. *Genome biology*, 11(11):R114.

Zhao, X., C. Li, J. G. Paez, K. Chin, P. A. Jänne, T.-H. Chen, L. Girard, J. Minna, D. Christiani, C. Leo, J. W. Gray, W. R. Sellers, and M. Meyerson

2004. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer research*, 64(9):3060–3071.

Zhao, X., B. A. Weir, T. LaFramboise, M. Lin, R. Beroukhim, L. Garraway, J. Beheshti, J. C. Lee, K. Naoki, W. G. Richards, D. Sugarbaker, F. Chen, M. A. Rubin, P. A. Jänne, L. Girard, J. Minna, D. Christiani, C. Li, W. R. Sellers, and M. Meyerson

2005. Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. *Cancer research*, 65(13):5561–5570.

Zheng, Z.-M.

2010. Viral oncogenes, noncoding rnas, and rna splicing in human tumor viruses. *International journal of biological sciences*, 6(7):730–755.